DTIC FILE COPY

AD-A212 879

Technical Report 851

Operator Workload: Comprehensive Review and Evaluation of Operator Workload Methodologies

Robert J. Lysaght, Susan G. Hill, and A.O. Dick
Analytics, Inc.

Brian D. Plamondon and Paul M. Linton
Sikorsky Aircraft

Walter W. Wierwille, Allen L. Zaklad, Alvah C. Bittner, Jr., and Robert J. Wherry
Analytics, Inc.

June 1989

DTIC
ELECTE
SEP 22 1989
S E D

United States Army Research Institute
for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

89 9 22 011

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS
--

2a. SECURITY CLASSIFICATION AUTHORITY
--

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE
--

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release;
distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Analytics Technical Report 2075-3

5. MONITORING ORGANIZATION REPORT NUMBER(S)
ARI Technical Report 851

6a. NAME OF PERFORMING ORGANIZATION
Analytics, Incorporated

6b. OFFICE SYMBOL *(If applicable)*
--

7a. NAME OF MONITORING ORGANIZATION
U.S. Army Research Institute
Field Unit at Fort Bliss, Texas

6c. ADDRESS *(City, State, and ZIP Code)*
2500 Maryland Road
Willow Grove, PA 19090

7b. ADDRESS *(City, State, and ZIP Code)*
P.O. Box 6057
Fort Bliss, TX 79906-0057

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
U.S. Army Research Institute for the Behavioral and Social Sciences

8b. OFFICE SYMBOL *(If applicable)*
PERI-S

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
MDA903-86-C-0384

8c. ADDRESS *(City, State, and ZIP Code)*
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
62785A	790	1102 (113)	C1

11. TITLE *(Include Security Classification)*
Operator Workload: Comprehensive Review and Evaluation of Operator Workload Methodologies

12. PERSONAL AUTHOR(S) Lysaght, Robert J.; Hill, Susan G.; Dick, A.O.(Analytics, Inc.); Plamondon, Brian D.; Linton, Paul M. (Sikorsky Aircraft); Wierwille, Walter W.; (Continued)

13a. TYPE OF REPORT
Interim

13b. TIME COVERED
FROM 86/09 TO 83/09

14. DATE OF REPORT *(Year, Month, Day)*
1989, June

15. PAGE COUNT

16. SUPPLEMENTARY NOTATION
Richard E. Christ, Contracting Officer's Representative.

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Workload, Man-machine relations
Human performance, Measurement, (SDW)
MANPRINT (Manpower and Personnel Integration),

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This report documents the results of an analysis of the scientific literature on operator workload. It begins with an extensive discussion of the concept and definitions of operator workload. The main body of the report is a review and analysis of techniques that have been used for assessing operator workload. These techniques are classified into two broad categories: (a) analytical or predictive techniques that may be applied early in system design, and (b) empirical or evaluative techniques that must be obtained with an operator-in-the-loop during simulator, prototype, or system evaluations. Information from the review provides practical guidance for selecting the most appropriate techniques for various system and resource characteristics. Keywords:

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Richard E. Christ

22b. TELEPHONE *(Include Area Code)*
(915) 568-4491

22c. OFFICE SYMBOL
PERI-SB

DD Form 1473, JUN 86

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Research accomplished under contract for the Department of the Army

Analytics, Inc.

Technical review by

Richard Dunn
Randall L. Harris
Sandra G. Hart
Thomas R. Metzler
Gary B. Reid
Jay Shively
Raymond Sidorsky
Michael Vidulich

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification______	
By______	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

NOTICES

~~DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600~~

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

ARI Technical Report 851

12. PERSONAL AUTHOR(S) (Continued)

Zaklad, Allen L.; Bittner, Alvah C., Jr.; and Wherry, Robert J. (Analytics, Inc.)

Technical Report 851

Operator Workload: Comprehensive Review and Evaluation of Operator Workload Methodologies

Robert J. Lysaght, Susan G. Hill, and A.O. Dick
Analytics, Inc.

Brian D. Plamondon and Paul M. Linton
Sikorsky Aircraft

Walter W. Wierwille, Allen L. Zaklad, Alvah C. Bittner, Jr., and Robert J. Wherry
Analytics, Inc.

Field Unit at Fort Bliss, Texas
Michael H. Strub, Chief

Systems Research Laboratory
Robin L. Keesee, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

June 1989

Army Project Number
2Q262785A790

Human Performance Effectiveness
and Simulation

The Systems Research Laboratory of the U.S. Army Research Institute for the Behavioral ~~and Social~~ Sciences (ARI) supports the Army with research and development on manpower, personnel, training, and human performance issues as they affect the development, acquisition, and operational performance of Army systems and the combat readiness and effectiveness of Army units. One concern that underlies all of these issues is the mental workload imposed on and experienced by the operators of newly emerging, high-technology systems, and the impact of that workload on operator and system performance. The Fort Bliss Field Unit is conducting exploratory development research for establishment of an operator workload assessment program for the Army.

This technical report provides a rigorous review of definitions of the concept of operator workload and of techniques that have been developed to assess operator and crew workload. Information from the review is integrated into preliminary guidelines for the selection and use of the most appropriate battery of techniques for each set of circumstances.

EDGAR M. JOHNSON
Technical Director

Preceding Page Blank

ACKNOWLEDGMENTS

Prudens interrogatio quasi dimidium scientiae

(Judicious questioning is virtually the half of knowledge.)

(Anonymous motto)

This project has derived leadership, guidance, inspiration, and support from Richard E. Christ and his colleagues at the Army Research Institute for the Behavioral and Social Sciences (ARI).

Special thanks are due to Sandra G. Hart of NASA-Ames Research Center, Michael Biferno of Douglas Aircraft Company, and Gary Reid and Glenn Wilson at AFAMRL, Wright-Patterson Air Force Base, for taking time out of their busy schedules to talk with us about workload. Barry Kantowitz offered valuable discussion about secondary tasks. Richard Monty and his group at the Human Engineering Laboratory filled us in on research activities. It is for these people that we have included the motto above: They helped us ask some of the right questions.

We owe a debt of gratitude to a TriServices Evaluation Panel organized by Dr. Christ for taking time to review and offer comments on an earlier version of this report. Their many comments improved this report immeasurably. Helene Iavecchia read the entire report and offered numerous comments about content and style. Nora Harrington used her blue pencil extensively on a number of chapters. We thank David Yensan for his comprehensive review for consistency.

Thanks to Leslie G. Tudor who reviewed literature and tackled the development of the secondary task tables. John Bulger provided many useful and practical comments on the manuscript. Scott Finkelstein and Eric Herman did many of the figures for the draft. Pam Lawfer and Elaine Kramer assisted in proofreading and formatting the text and in lots of other ways. Administrative tasks were ably handled, as usual, by Heidi Brauer-Langdon.

And, finally, many thanks to our colleagues at Analytics, Inc., for their patience while we worked long hours, hogged the computers, and complained about our workload.

OPERATOR WORKLOAD: COMPREHENSIVE REVIEW AND EVALUATION OF OPERATOR WORKLOAD METHODOLOGIES

EXECUTIVE SUMMARY

Requirement:

The overall purpose of this report is to provide useful and practical information concerning operator workload (OWL). It is specifically aimed at information applicable in conceptualizing, specifying, designing, developing or evaluating systems for the Army.

Procedure:

Relevant research and published materials were identified and obtained through libraries and personal contact. The literature obtained was reviewed for specific OWL techniques. A workload technique taxonomy served as the organizational scheme within which the workload literature was reviewed. Organizations engaged in significant workload research were visited to discuss current and ongoing OWL research.

Findings:

Operator Workload is explained and defined using several informal examples, definitions of workload used by researchers as reported in the literature, the foundation of a general definition, a framework (taxonomy) to organize the various workload estimation techniques, and some general issues concerning the techniques. After considering a variety of performance issues and definitions in the literature, the idea of a performance envelope is developed. Workload determines the current position in the envelope. The primary interest is the operator's position relative to the boundaries of the envelope and the operator's relative capacity to respond.

Techniques that have been used for assessing OWL and determining the operator's current and future position in the performance envelope are reviewed and analyzed. These techniques are classified into two broad categories:

- Analytical--predictive techniques that may be applied early in system design without an operator-in-the-loop, and
- Empirical--operator workload assessments that are taken with an operator-in-the-loop during simulator, prototype, or system evaluations.

The analytical techniques can be used early in system design when there is greatest design flexibility and throughout the materiel acquisition process. The analytical category includes comparison techniques, mathematical models, expert opinion, task analyses, and simulation. Considerable progress has been made in developing workable analytical tools but much remains to be done.

Empirical techniques are used when operators and a simulator, a prototype, or a system are available for testing. The empirical category includes primary task measures, subjective methods, secondary task techniques, and physiological techniques. Each of these subcategories is discussed in separate chapters. Descriptions of the methods and techniques are provided, along with discussion concerning available information about their validity, reliability, sensitivity, diagnosticity, intrusiveness, and practicality. Recommendations for application are included with the discussion of individual techniques.

Utilization of Findings:

The information from the reviews is integrated into a foundation for a practical guide for the user. Example case studies are provided, along with suggestions for the most appropriate techniques, both analytical and empirical, to use for various system-resource characteristics. A working guide is also provided for a general approach to the selection and application of workload techniques. This application guide encompasses all major issues. Twenty-three general questions are developed to assist in identifying the proper techniques. The answers provided by the user aid in the selection and application of techniques. These include general questions about stage of system development, category of system, and resources available--both personnel and equipment, and a number of specific questions about workload.

OPERATOR WORKLOAD: COMPREHENSIVE REVIEW AND EVALUATION OF OPERATOR WORKLOAD METHODOLOGIES

CONTENTS

CONTENTS (Continued)

LIST OF TABLES

CONTENTS (Continued)

LIST OF FIGURES

Operator Workload: Comprehensive Review and Evaluation of Operator Workload Methodologies

CHAPTER 1. INTRODUCTION

"The human factors in most practical situations have been neglected largely because of consciousness of ignorance and our inability to control them [human factors]. Whereas engineers deal constantly with physical problems of quality, capacity, stress, and strain, they have tended to think of problems of human conduct and experience as unsolved or insoluble. At the same time there has existed a growing consciousness of the practical significance of these human factors and of the importance of such systematic research as shall extend our knowledge of them and increase our directive power.

"The great war from which we are now emerging into a civilization in many respects new has already worked marvelous changes in our points of view, our expectations and practical demands. Never before in the history of civilization was brain, as contrasted with brawn, so important; never before, the proper placement and utilization of brain power so essential to success.

"Reprinted in part from a Harvey lecture delivered by Major Robert M. Yerkes in New York, January 25, 1919, and published with the approval of the Surgeon General of the Army, from the Section of Psychology of the Medical Department." In turn, reprinted from: Yoakum, C. S. & Yerkes, R. M. (Eds.) (1920). *Army Mental Tests*. New York: Henry Holt and Company.

There are several noteworthy points about this quote from 1919. First, many problems about "human conduct" can now be solved. Techniques have been developed in the last seven decades which are applicable to these problems, and furthermore, engineers are using the results of these techniques as design principles. This report is a testament to these techniques. However, just as evolving technology produces better and more sophisticated hardware, technology will evolve to produce even better and more sophisticated assessment techniques. Second, the trend for "brain, as contrasted with brawn," has accelerated. Indeed, the bulk of the workload literature deals with brain and not brawn.

The Changing Role of Army Operators

Technology is becoming increasingly advanced and complex. As new systems are developed, new technologies are employed, and the role of the operator is changed. The newest generation of advanced military systems uses advanced computer technology for multifunction displays, decision aids, intelligent systems, or computationally-assisted control. Technological advances have resulted in changes to operational procedures and the functions of the system operators. Operators perform more planning, supervisory, monitoring and overseeing functions than in the past. In many instances computers are doing the computational work and the operators are continually checking for system failures or emergency conditions. It seems fair to characterize the changes in operator functions as more mental or cognitive in nature. Furthermore, operators are often required to perform these functions in stressful and physically demanding environments.

A plausible scenario could have an operator sitting in front of one or more computer displays. The displays contain information which must be processed and acted upon. Several potential targets are displayed and the operator must decide which, if any, should be fired upon and with what priority. In this scenario, the operator is one member of a crew who is expected to perform both night and day, even when fatigued. Some functions can be shared among the crew members, others can not. The amount and rate of displayed information is high; communications channels are open and busy; decisions must be made within seconds. In this situation, the single operator or crew may not be able to perform the required tasks within the critical time window. This situation may lead to operator overload resulting in performance degradation *AND* mission failure. This generic scenario is applicable to many emerging combat systems and this report is concerned with one specific part of this problem: *Operator Workload (OWL).*

Current Status of OWL in the Army

MANPRINT is an Army initiative which considers the role of the soldier in system performance. Through this initiative, the Army addresses the question, Can this soldier, with this training, perform these tasks to these standards under these conditions? The Army MANPRINT guidance is contained in Army Regulation (AR) 602-2, *Manpower and Personnel Integration (MANPRINT) in the Materiel Acquisition Process* (U.S. Army, 1987). It is clear that AR 602-2 requires MANPRINT issues be addressed, and hence that human performance data be obtained and analyzed at all stages of the Materiel Acquisition Process (MAP). It is also increasingly apparent that the MAP often does not allow consideration for the possible effects of exceeding OWL capacity. Due to changes in technology, cognitive overload is more likely than in the past and this cognitive overload can easily induce operator errors and cause critical information to be processed incorrectly or missed entirely, leading to a degradation of system performance.

While it can be argued that OWL concerns are not synonymous with MANPRINT, OWL is related to the six MANPRINT domains. These domains are:

- Manpower
- Personnel
- Training
- Human Factors Engineering
- Safety
- Health Hazards

Consideration of the interrelations between OWL and the MANPRINT domains will assist in identifying MANPRINT trade-offs that may be made in an effort to maximize system performance. For example, economic pressures to reduce crew sizes (manpower) has immediate impact on operator workload. As new devices are added to replace humans, the workload of the reduced crew certainly changes and the perceptual and mental workload of individual operators may actually increase. This, in turn, has impact on the interrelation between OWL and personnel issues which involves trade-offs between soldier quality (as measured by the Armed Services Vocational Aptitude Battery [ASVAB]) and the degree to which soldier perceptual, mental, and psychomotor loading occur. Further, workload may vary due to training, soldier quality, soldier-machine interface and the degree of soldier information loading. Knowledge of the OWL-related requirements may assist in better, more efficient personnel and training actions. The MANPRINT domains are overlapping, and because of this a change in one domain will have an influence on others including OWL. Clearly, MANPRINT and workload concerns are interrelated.

A requirement has been established that OWL issues need to be addressed at all stages of the MAP. The regulation AR 602-1, *Human Factors Engineering Program* (U.S. Army, 1983), specifies that the Human Factors Engineering (HFE) program shall be performed in accordance with MIL-H-46855B, *Military Specification: Human Engineering Requirements for Military Systems, Equipment and Facilities* (U.S. Army, 1979). This latter military specification (Section 3.2.1.3.3) requires that individual and crew workload analyses shall be performed and compared with performance criteria. However, no guidance is provided to the system developer as to how such a workload analysis should be performed (Hill & Bulger, 1988). This lack of guidance has led to the effort which comprises the body of this report.

Purpose of the Report

A goal of this report is to present a review of currently available methodologies and techniques that have been developed and used in the assessment of OWL. In this effort, more than 1500 reports were reviewed and close to 500 research reports are cited. This review was intended as a critique of the methods and techniques that have previously been used to examine workload. It contains descriptions of the methodologies and techniques as well as discussions concerning the available information regarding validity, reliability, sensitivity, intrusiveness, and practicality. In addition to methods and techniques that have previously been used to assess workload, other methods are also identified that may be applicable to OWL.

A second equally important goal of this report is to analyze and integrate these methodologies into a practical guide for the user. Thus, the reviewed techniques are analyzed with respect to reported effectiveness and resources needed for implementation. References guide the reader to sources for additional information. In addition, sample applications are considered in Chapter 8. The overall purpose of this report is to provide useful and practical information concerning OWL to those involved in conceptualizing, specifying, designing, developing, and evaluating systems for the Army.

Methodology Used in this Report

The approach of this comprehensive review of OWL research and methodologies had two major thrusts. The first was to provide a technical review and analysis of available literature related to OWL. The second thrust was to be aware of the practical utility and importance of OWL issues to the Army. In recent years, OWL has received considerable research attention reflecting its importance, and efforts continue to understand theoretical as well as application issues. The practical ways in which workload issues could impact system performance, conceptualization, design, development, and evaluation were considered at all times.

Review Approach

Relevant research and published materials were identified and obtained through libraries and personal contact. The literature obtained was reviewed for specific OWL techniques. The workload technique taxonomy, described in Chapter 2, served as the organizational scheme within which the workload literature was reviewed. Organizations engaged in significant workload research were visited to discuss current and on-going OWL research. These included Douglas Aircraft Company; NASA-Ames Research Center; Wright-Patterson Air Force Base; USAHEL; and NASA-Langley Research Center.

The usefulness of the various techniques for addressing Army needs was the focus of the project. Particular emphasis was placed on the sensitivity of the OWL techniques for measuring differences in various tasks. In addition, other important practical criteria that received particular consideration are the intrusiveness of the techniques and the relative costs and the level of expertise needed for their use. Descriptions of the techniques and discussion concerning their implementation in Army applications are given throughout the volume.

Organization of the Report

This report presents a review and synthesis of literature related to operator workload. Each chapter begins with a brief discussion of the purpose of the techniques. In the body of the chapter, definitions, details, and examples of the techniques are given as well as research concerning the specific techniques . Included in these discussions are comments about issues concerning salient characteristics which will be defined later; these include the issues of sensitivity, diagnosticity, intrusion, validity, reliability, implementation, operator acceptance, and relative cost of use of key techniques. These criteria were chosen as important to practitioners and as appropriate to characterize the methods.

In Chapter 2, basic issues concerning OWL are discussed, including the definition of operator workload, a taxonomy of workload assessment methods and techniques, as well as other important general OWL issues. Subsequently, the descriptions and discussions related to specific workload techniques and methodologies are presented in Chapter 3 for Analytical Techniques and Chapters 4 through 7 for Empirical Techniques. The organization of these five chapters follows the organization of the taxonomy to be presented in Chapter 2. Chapter 8 describes an approach for the selection of appropriate techniques for assessing workload. Finally, a concluding and summary Chapter is provided including a different perspective on OWL and indications of some future directions.

CHAPTER 2. THE CONCEPT OF WORKLOAD

What Is Workload?

This chapter provides a discussion of the concept of workload and some definitions. It provides the formal background for the reviews and evaluations of the specific workload assessment techniques discussed in subsequent chapters. However, in order for the reader to have an intuitive understanding of operator workload, several examples are presented first.

An Example: Driving

As an illustration of what is meant by the term workload, imagine you are driving in your favorite car. As you go through this mental exercise, we will increase the difficulty with each successive statement in a number of different ways. Additionally, we will use some words like stress and effort in a colloquial manner; these will be defined more precisely later. When we are through with the exercise you may not know exactly what workload is, but you will have a feel for the range of operator workload possible for a task as common as driving a car. And more importantly, you will have a feel for the importance of workload and the various factors that affect workload. One point we wish to make in this example is that workload is not the same as performance.

- Since you are an important personage, the State Police have closed the Interstate to all other drivers. You are cruising down the highway at the speed limit on a nice, sunny day . Easy driving, right?
- You have just passed the state line. This second state doesn't think you are quite as important and now you have some traffic. Still not bad.
- You have been driving for a while, it is approaching the rush hour near a metropolitan area and traffic is picking up.
- It is Friday afternoon and every one wants to get home or out of town before the storm hits. Traffic is now much heavier than normal and slowing down. (We must be in Connecticut.)
- You left early this morning and didn't realize you hadn't stopped for lunch. You're tired and hungry.
- Traffic is now reduced to a crawl. You also forgot to get gas when you forgot lunch. You've got to get to an exit and find a gas station.
- While you are crawling along, the weather has turned. It is now raining.

Preceding Page Blank

- It has also gotten dark and visibility is not good. The highway is not well marked and you must be careful not to miss your turnoff.
- Worse, the car in front does not have brake lights so you have to pay very close attention to this stop-and-go stuff. Eyes on car in front.
- A few miles are covered, but with the dark, the outside ιemperature has also dropped. It is no longer just raining, it is freezing. Several cars are off the road. Still bumper to bumpe" and gas is getting very low.
- Your two year old, who was sleeping in the backseat, wakes up. He is hungry, scared, and crying.
- It's not a lot of fun with all that is going on. In addition, the engine sounds like it is missing and you know you are not yet quite out of gas. (You've turned the radio down and would like to turn the kid down.)

You are about to 'lose it' as anyone who has been in a similar situation can attest. Improbable, yes, but not impossible. (And note, we didn't cheat by giving you an unfamiliar vehicle with shift instead of automatic, or even an English car with the wheel on the 'wrong side.' We assumed that your prior training and experience was in effect.) Further, we didn't even have hostiles shooting at you. Nor did we have you crash - - **Performance remained acceptable.**

A Second Example: Mental Load

Before we start discussing workload in a formal way, we want to consider one more example, this time strictly mental load. First, we are going to ask you to do a couple of tasks that are highly overlearned and very easy. Then we will do the tasks again, but in a combined manner. Not only does the demonstration illustrate an example of cognitive workload, it illustrates an important point about measuring workload: Two easy tasks added together can sometimes result in a very difficult task. Not an easy situation to predict. As you do the task, take your time. You might even want to time yourself on each of the parts.

- Recite the alphabet,
- Count from 1 to 26,
- Now do both, interleaving the alphabet with the counting, A-1, B-2, etc. saying the answers.

If you actually got all the way through the combined task, you are unusual. Most people give up about G-7 or H-8. Why is it so difficult? Let us use this example to diagnose the basis of the difficulty and illustrate workload analysis. Get out a pencil and a piece of paper. Do the double task again, this time writing down the answers. Any difficulty in getting all the way through this time? Part of the difference between the two is that the pencil and paper reduces the heavy burden on memory. There are some additional reasons, but the point is that the same task can be difficult or relatively easy depending on how we do it. And we can identify the reasons for the differences. In this example, there is usually a performance failure on the first attempt which burdens memory and success on the second attempt which uses pencil and paper - - **Performance is acceptable only in the second case.**

These two examples should give you an idea about the variation of difficulty of tasks and the difference between measuring performance and the amount of effort you have to expend to perform the task.

In this chapter, we consider a number of general issues involving workload. Later chapters will cover more specific, detailed issues. First we present a description of the relation of performance and workload. This leads to a discussion of some human performance concepts in the context of system operation. To a large degree, this discussion is the foundation of all that is to come later. Then, the chapter provides a review and discussion of definitions of OWL. Also included is an organization of workload assessment techniques in the form of a taxonomy that provides a structure within which to classify the measures.

Performance vs. Workload

Performance is what we are ultimately concerned with, Can the operator successfully complete the mission? One goal of workload research is to predict impending doom - failure of performance. Not only do we *not* want the mission to fail, we also do *not* want the man or machine to be damaged. Having anticipated and predicted a trouble spot, the second goal is to correct those situations in which performance fails. As an aid in this effort toward better and safer performance, researchers have developed the concept of workload.

The relation between workload and performance is illustrated in Figure 2-1. In the figure, it can be seen that workload and performance seem to have an inverted U relation. At extremely low levels of workload as in Region 1, the operator may become bored (Hart, 1986a). Boredom can lead to missed signals and instructions, resulting in poor performance (Parasuraman, 1986). (Although this report will not address cases subsumed in Region 1, it is well to note that performance can be adversely affected if OWL is too low as well as too high.) With a reasonable level of workload, performance can be expected to be acceptable as shown in Region 2. However, further increases of workload into Region 3 show a marked degradation in performance. Figure 2-1 also illustrates that workload is not the same as performance. Performance may remain at an acceptable level over a considerable range of workload variation as in the driving example. In general, however, workload extremes are related to poor performance.

Internal and External Performance Limitations

There is general agreement about many of the determinants of good or poor operator performance. Norman and Bobrow (1975), for example, differentiate between two categories of limitations on performance: data-limited and resource-limited. Data limitations occur when task processing is constrained by unavailable data, e.g., trying to read a map in the dark. This is a limitation external to the operator; stimuli may be below threshold or may contain insufficient information to solve the problem. By contrast, resource limitations occur when the human information processing system cannot handle the data rapidly enough. In this case, performance decrements are due to internal limitations. In either case, performance decrements can be observed in several forms, gradual, intermittent, or catastrophic. One of the goals of OWL research is to uncover, identify, and eliminate those instances in which the demands of human tasks would degrade human and system performance.

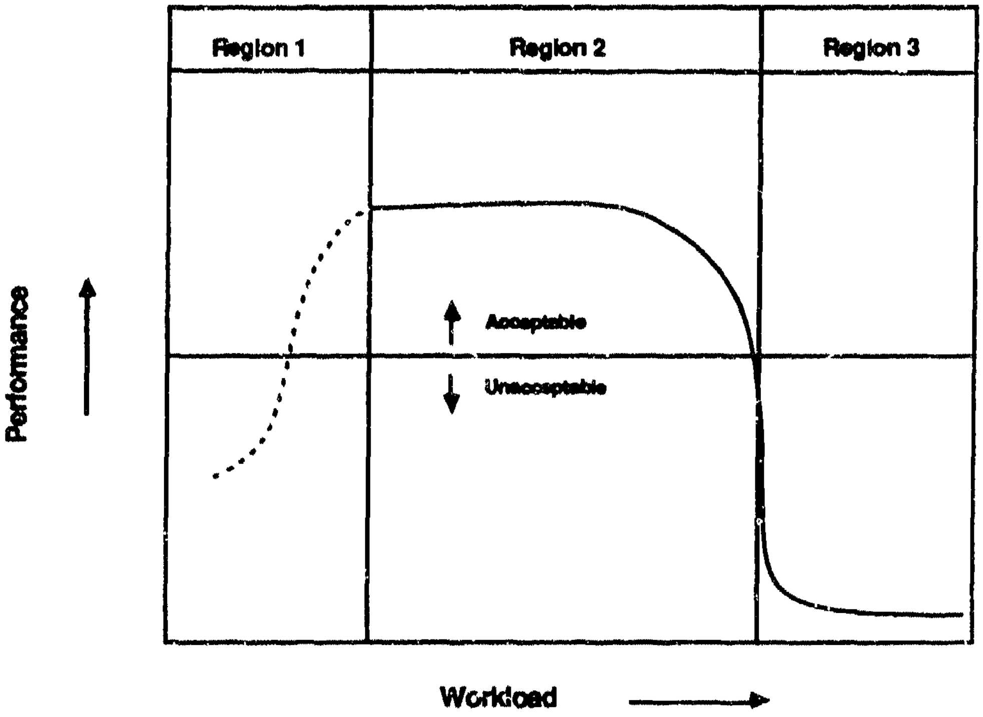


Figure 2-1. The hypothetical relationship between workload and performance. (This figure is a compilation of the concept discussed in several places [e.g., Hart, 1986a; O'Donnell & Eggemeier, 1986; Tole, Stephens, Harris, & Ephrath, 1982]).

Presumably, the data-limited decrements should be eliminated in the design phase of a system. Dials and gauges should be easy to read; communications should be easy to understand, and all key data accessible. However, to the extent that necessary information is simply not available during a mission, the operator must seek the information from other sources or spend additional time estimating parameters needed for decision making. This illustrates the important point that the operator, the system hardware, and the environment all interact in affecting performance and this interaction can change the nature of the task. The form of the interaction can also have important consequences for mission performance.

A Model of the OWL Context: Factors Affecting Performance and OWL

The previous discussion illustrated one way of looking at performance limitations and interaction of the human with the environment. Because human behavior is dynamic, such interactions abound - - much to the frustration of the workload researcher. To help the reader understand the intricacies of behavior, performance, and workload, a brief discussion of the variety of influences on the operator is presented.

Performance is affected by two major kinds of factors: (a) the operator tasks defined by the mission, by the environment, and by the design of the workstation and (b) the transitory states and stable traits of the human operator. Figure 2-2 illustrates these factors, all of which combine to influence how the individual will respond to the ongoing demands. The interaction of these factors will determine both operator workload and operator performance and, hence, system and mission performance. Each of these components is considered in more detail below. The upper portion of the figure contains some external influences. The system design, mission and other external factors combine to create situational demands for the operator. In the middle of the figure is represented the operator including a breakdown of some of the internal factors of the operator which have a bearing on OWL. At the bottom of the figure, the ovals represent approaches to obtaining responses from the operator which are used to make inferences about the operator. It is important to note that the bottom oval, system performance, is directly related to MANPRINT concerns. More will be said about these measurements in Chapter 4.

Situation Demands and External Influences

Mission Requirements and Task Allocation. The allocation of system functions to the human is an initial step in system design and this allocation will, in turn, lead to situation demands on the operator. During system design, the design team decides which functions are allocated to humans and which are allocated to the system. Once allocated, those functions plus the design of the controls and displays will define the operator tasks. The tasks allocated to a given operator represent that operator's job. The

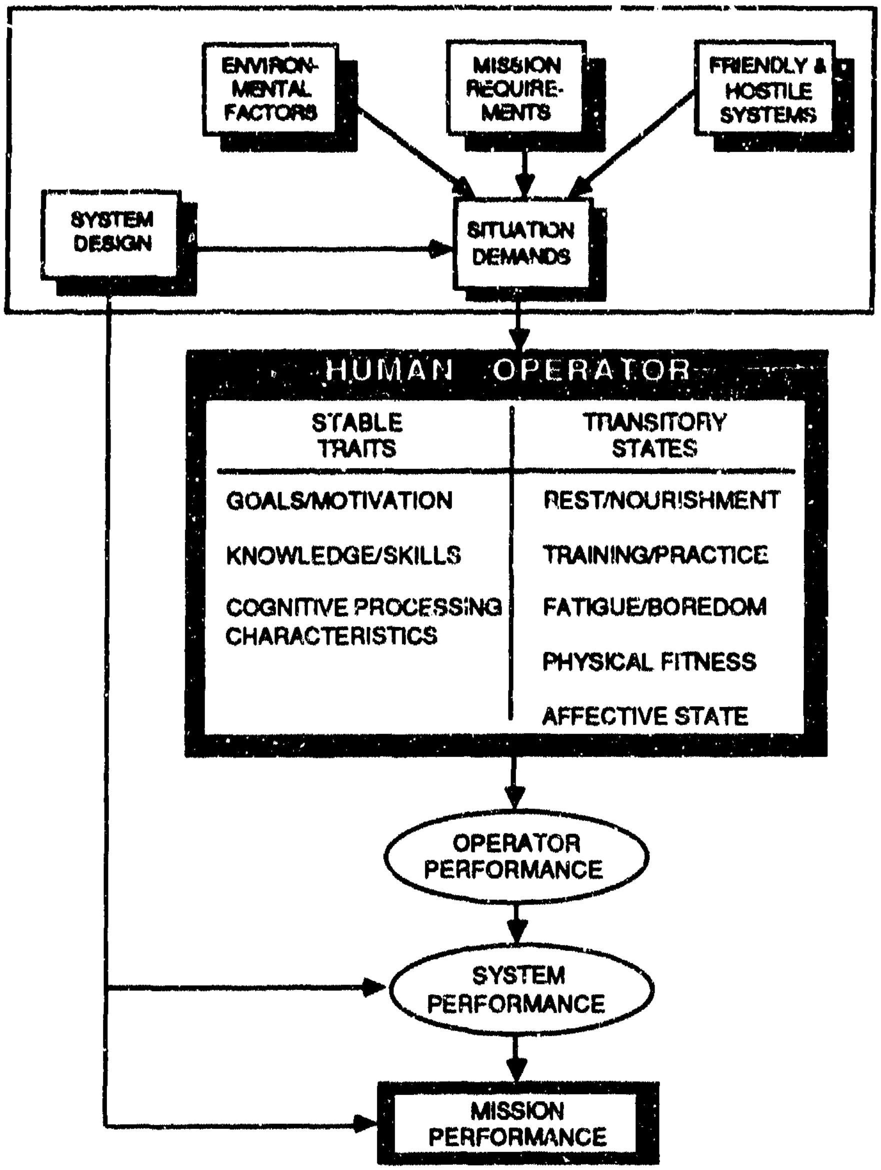


Figure 2-2. A conceptual framework of the OWL context and influences on operator/system performance.

human factors technique of task analysis is concerned with understanding how these tasks will impact the probable overall performance of the operator, and the extent to which some of these tasks might not be performed at acceptable levels.

Two tasks may differ in a variety of ways which can affect their accomplishment. The two tasks may require different types of actions. In turn, those actions may require more effort or time by the operator than does another task. Regardless of the type of task, the operator must perform some sequence of acts on some objects or entities in order for the task to be accomplished. In some tasks, a majority of the required actions involve manipulations of physical objects. Other tasks may be dominated by actions requiring the operator to sense or perceive the attributes and characteristics of objects. Still other tasks exist in which a majority of the actions involve manipulations of internalized definitions, facts, or concepts.

Similarly, two jobs may differ in the kinds of tasks required by an operator, and in the sequence in which the tasks must be performed. Some jobs may have many tasks that do not overlap in time. Other jobs may have multiple ongoing tasks during the same time periods and require the operator to time-share among those tasks.

Finally, the system's machine capabilities (e.g., sensor, data processor, and propulsion subsystems), the relative capabilities of hostile forces, and the availability and capabilities of cooperating, friendly forces will change from mission to mission, and will impact the speeds and accuracies with which various operator tasks must be accomplished.

In summary, tasks can influence the workload that will be imposed on the operator by:

- Actions required by each task,
- Sequence of actions performed for a task,
- Number and types of tasks to be performed,
- Time available for each task to be completed,
- Overall time constraints, and
- Required accuracy levels.

Taken together, these influences constitute a comprehensive set of factors that contribute to the situation demands illustrated in Figure 2-2.

The Environmental Context. The tasks performed by the operator are not done in isolation, however. A given task may occur in widely differing circumstances that can affect the level of difficulty of that task for the operator. The way in which the operator interacts with the immediate surroundings will also have important implications for performance and workload. It is widely recognized by engineers that machine components cannot tolerate some kinds of physical disturbances. They must be protected (hardened) to function in the presence of hostile environments. Detailed attention is given to specifying how machine components will be packaged, supported, and interfaced with other machine components. Similar attention must be given to the support and interfacing of humans, both with one another and with machine components. Among the external factors which alter situational demands and which affect levels of task difficulty are:

- The external environment in which the task must be performed (e.g., heat, humidity, sound, illumination, vibration, and g-forces)
- The design of the human-machine information exchange units (e.g., types and sizes of displays and controls, and their layouts and formats)
- The design for human packaging (e.g., protective clothing, seating, and restraints)
- The design of the overall workstation (e.g., its size, internal lighting, ventilation, temperature and humidity control, and vibration dampening)

To a large extent, external environmental factors cannot be controlled by the system design team; these are determined by the missions. However, the immediate external environment and the extent to which it impinges on the operator can be partially controlled by other design factors. Because many operator tasks involve the exchange of information between the machine and human, the design of the operator console will affect human performance on the tasks. Both the speed and accuracy with which the operator can perform a given task and the extent to which the operator can maintain acceptable performance for long periods of time will be partially dependent on the ambient environment. Thus, operator support and workstation design factors will influence the workload of the operator.

The Operator

Every operator enters into a situation carrying a number of influences which can impact performance. These are divided into transitory which can be modified relatively easily and stable which are much more difficult to modify.

Transitory States. Transitory states can be considered to be initial states such as the amount of rest level of physical fitness, etc. which may or may not be appropriate for the mission. These are depicted in the center right potion of Figure 2-2. Training is, of course, an important factor. Indeed, training is sometimes considered to be the single most important factor in mission success/failure and often a panacea: If the mission fails, provide more training. Certainly, training and specific skill acquisition are important and extend the operator's capability to handle workload (Bainbridge, 1978). In the context of Figure 2-1, this would be represented by increasing the effective area of Region 2. Harris, Tole, Stephens, and Ephrath (1982) have expressed similar ideas. However, there are numerous aspects of high workload which cannot be handled by additional training, for example, the requirement to perceive faster. Many of these high workload factors are related to the cognitive processes of the operator.

Stable Traits. In addition to transitory states, the human operator is characterized in the left center portion of Figure 2-2 by several interrelated facets which change slowly over time: goals/ motivational state, knowledge/skills, and processing capabilities. Processing capabilities refer to the operator's higher-level behavioral components (e.g., thinking) which interacts with and integrates knowledge and skills to accomplish task element goals.

Individuals may differ in the relative importance of various goals, the extent to which those goals are currently satisfied, and the extent to which performing a given task is perceived as being important to goal achievement. They may also differ in their perceptions of the speed and accuracy with which a task needs to be done. These factors, in turn, determine the level of motivation for task accomplishment and, consequently, the effort an individual is willing and able to put forth in accomplishing the task. The motivational aspect of the workload often is ignored by researchers. Gopher and Donchin (1986) handle the motivation issue by ruling it out; they assume that every operator is highly motivated and wants to maximize his or her performance.

The cognitive processing capabilities of an individual are distinguished here from the knowledge and skills an individual has acquired through training and experience. Knowledge (e.g., facts, rules, equipment usage procedures) can be considered as a resource of the individual to be utilized by cognitive processes. To use that knowledge, however, the individual must invoke other dynamic processes to retrieve and manipulate the knowledge required to execute a task. Other cognitive processing capabilities are needed to glean information from displays and to manipulate controls.

Individual Differences among Operators

Humans are known to differ in terms of individual traits or capacities that can impact task performance. Two individuals may differ from each other in a variety of ways which may make accomplishment of the same task easier, faster, or better for one individual than for the other. Physical size and strength are two obvious dimensions along which differences may be observed.

More important in modern technological systems are the mental and cognitive differences among individuals. A list of the important cognitive components is probably longer than a list of the researchers studying the problem. Some of these variables include information processing, perceptual processing, decision making, numerical operations, and spatial processes used for tasks such as map reading. Some of these variables are represented in the Armed Services Vocational Aptitude Battery (ASVAB). The aviation community has led the way in using such tests in selection. However, it is probably fair to say that little research has been done exploring individual differences in cognitive skills in the context of workload.

Summary

This, then, is the situation we need to study and it is complex. Clearly, workload and human performance are affected by external influences, and operator states, both transitory and stable. How do we measure performance success or failure? Because there are many determinants of performance, researchers have devised many ways to predict and to measure their influence on behavior. Each method may provide different answers. Thus, the way the question is phrased and the approach to assessing behavior becomes important. Consequently, workload is generally considered to be complex and a multidimensional concept.

Definitions of Operator Workload

A parade of definitions can be rather dull. However, various authors have discussed the meanings and definitions of workload in different manners. Even though no single definition of workload is generally accepted, it is well to organize the various threads of thought into a more coherent and practical package. Accordingly, a review of the alternative definitions will be instructive. What we will find is that each author has a different twist and this twist is reflected in associated research efforts. The differences often stem from an incomplete understanding of underlying mechanisms and processes. So it is in workload; workload is not a unitary concept but, in fact, a multidimensional one. The particular definition one adopts has extremely important implications in the application of the various techniques to measuring workload.

Webster's defines workload in the following ways:

> **workload** *n* 1: amount of work or of working time expected from or assigned to an employee. 2: the total amount of work to be performed by a department or other group of workers in a period of time (Webster's Third International Dictionary, 1976, p. 2635).

A scientific definition becomes much more detailed than just **amount of work or of working time**. Rather than just considering the individual, one can consider parts of the individual. Thus, one can analyze the amount of work done by the hands or by the eyes, or any other part of the body. A common distinction made along these lines is between physical and mental workload. Similarly, the definition implies some external agency defining the amount of work and the number of things to be done. Bosses are good at that. However, for purposes of argument, we could also consider workload from the employee's viewpoint. Comparing the two viewpoints may show a discrepancy! Indeed, we will discuss the viewpoint of some investigators who state the latter viewpoint is the correct viewpoint.

Webster's second definition refers to crew workload and will not be discussed in this volume. Individual operator workload relates to personnel and training considerations; crew workload relates to manpower considerations as well. At a basic level, the term workload carries a number of meanings within the military community, especially the second dictionary definition. In particular, within a MANPRINT context, workload often is associated with the number, frequency and durations of activity-based tasks performed by a specific number of Army personnel of particular Military Occupational Specialities (MOS's), skill levels, and paygrades. It is clear in this context that workload does not refer to cognitive/physical underload or overload, but rather to task-based manning considerations. Obviously, care must be taken to specify clearly what is being discussed when using terms like workload and workload analysis. Crew workload and manpower considerations are closely tied to the potential cognitive overload of individual operators (Hill & Bulger, 1988), but they are different and should be clearly differentiated.

The discussion of the human operator model suggests that an operator's performance on a given task depends not only on the demands of the task, both in accuracy and time, and the situation in which it is embedded, but also on the capability and the willingness of the operator to respond to those demands. A difficulty in defining operator workload is that there are alternate, legitimate ways in which workload can be considered. We will not consider all possible definitions, but rather just the set that has been most often used by the researchers. To a large extent, definitions depend on the techniques used and the constraints imposed by those techniques. In this section, three broad categories of workload definitions are discussed:

- amount of work and number of things to do,
- time and the particular aspect of time one is concerned with, and
- the subjective psychological experiences of the human operator.

We will consider each of these categories from several vantage points. The first two are congruent with the first dictionary definition and have parallels with traditional time and accuracy performance measurement. The psychological dimension is added. Doing so reveals gaps in research which obviously have implications for application. Although it is somewhat premature, we will also relate the definitions to follow to the workload assessment techniques employed.

Every reader is familiar with the fable of the three blind men examining the elephant. Each of the blind men was right is his observation but wrong in his conclusion. Much of what will be discussed in the next section is a living example of this fable. But science is like that. We obtain one observation at a time, and through a collection of observations, the truth begins to emerge. Later we will describe the elephant called workload. First, however, let us review some observations.

Amount of Work / Number of Things To Do

To quantify operator workload, some researchers have sought to identify the absolute amount of work required to complete a given task. Although this is a desirable goal, it must be recognized that the actual amount of work needed to complete a given task (e.g., assessing a tactical situation) varies with the situation (e.g., the number of targets on a screen). Nevertheless, a given task will still possess a distribution of amounts of work required, and it might be useful to estimate that distribution. This approach to quantifying workload considers it as a function of the task and situation – a point of view external to the human operator. A parallel conception is the number of things which have to be done in a psychomotor context (Dick, Brown, & Bailey, 1976). Both of these conceptions are performance based. Note that this conception implies an accuracy or a quality component of human performance. The quality is not always well defined; sometimes it is just in terms of satisfactory completion: 'Any landing you can walk away from is a good landing.'

The concept of work in the physical sciences is readily understood. It is sometimes less clear what work means for biological systems. There is a large overlap in the concept of work for machines and humans, and it is instructive to describe an analogy between them. First, work is not performed without some cost. Energy or other resources must be expended for work to be accomplished, for example, gasoline is stored in a vehicle's tanks, electricity is stored in batteries, etc. Second, the burning of fuel and oxygen results in energy being released. Third, the rate at which fuel is burnt may change from moment to moment depending on the current demands of the situation. The vehicle could also run out of fuel. Something or someone must detect or be aware of the changing situational demands and regulate the rate at which fuel and oxygen is being delivered to the engine.

The Stable Capacity of an Individual. While most would agree that the amount of work to be done is an important element of task workload, the amount of work must be considered in relation to the capacity of the individual to perform that work. Here again, there are excellent analogies to mechanical workloads. For example, we may define a task as moving a wagon having a particular load from one location to another. A vehicle having a large capacity motor may experience no difficulty in performing that task. However, as the capacity of the motors of alternative vehicles gets smaller and smaller, greater and greater difficulty will be experienced in performing that task. In fact, at some point, the load might be too much for one of the vehicles to handle. In the same fashion, humans differ in their capacities to perform a given task. Some might find a task easy to do while others might find that same task impossible to perform. This viewpoint of workload represents a conception of workload internal to the operator rather than external to him. Furthermore, the capacity of the human is assumed to be fairly stable across time, as what might be found by administering personnel selection tests.

There are two different meanings for the term 'capacity'. One involves considerations between individuals (individual differences) and the way performance and workload differ from one individual to another. Little work has been done in this area. The other meaning refers to a single individual and is used in the context: How much more can the operator do? This latter meaning has been considered in much greater detail by researchers .

Spare Capacity of an Individual to Perform Other Tasks. Much discussion of workload has been based on the foundation of information processing concepts. Gopher and Donchin (1986) suggest that workload implies "limitations on the capacity of an information processing system" (p. 41-3). Gopher and Donchin (1986) and Kantowitz (1985; 1987a) overview some of the more prominent theoretical models related to the workload area. In these overviews, a major theoretical perspective of workload is the spare capacity model. Under this formulation, the human is viewed as having a limited capacity or ability with which to process information. A simplistic example would be a person who has the capacity to receive and process a specific amount of information. If that person is currently using only 25% of that capacity, then the person has 75% spare capacity currently not in use.

Resources Available. A related model also based in information processing is referred to as the *multiple-resources theory*. In this theory, multiple pools related to specific abilities, such as verbal and spatial, are postulated to exist. Workload is then considered in the context of utilization of the abilities, singly and in combination. Much work has been done in support of this theory (e.g., Navon & Gopher, 1979; Wickens, 1980; 1984) that suggests there will be less competition for the limited resources, and hence less overall workload, when controls and displays do not all require the same resource pool (e.g., verbal) for processing and controlling than if the display and associated control require the same resources. (Our second example at the beginning of this chapter, interleaving recitation of the alphabet and numbers, is an example of competition for memory resources.) From this perspective, "mental workload can be described as the cost of performing one task in terms of a reduction in the capacity to perform additional tasks, given that the two tasks overlap in their resource demands" (Kramer, Sirevaag, & Braune, 1987, p. 146). However, this theoretical perspective has its skeptics who suggest that single pool capacity is sufficient; multiple pools of capacity are simply unnecessary to explain human information processing (e.g., Navon, 1984; Kantowitz, 1987a).

Time Based Conceptions - Working Time

The preceding section discussed several ways in which researchers have described workload in terms of amount. In this section, we consider the issue of time. Three different ways of considering operator workload are described, all based on temporal elements. Each defines workload in relation to some time component, as in the amount of something that has occurred, is occurring, or is scheduled to occur. Simply defining workload as amount of working time fails to inform us of whether we should attend to (a) the past, work completed, (b) the present, work currently being accomplished, or (c) the future, work scheduled and work anticipated.

The future is the easiest to deal with. To date, there have been few published discussions of work scheduled as a factor determining workload. Nevertheless, the current activity of an individual will be influenced by what has to be accomplished later. As Hart (personal communication, July 1987) has pointed out, the amount of time spent on a current task is influenced by the known and expected time requirements of future tasks. Sheridan and Simpson (1979) call this nearness to deadlines.

One of the most commonly used conceptualizations of time involves the present. It includes the time required (Tr) for a task in relation to the time available (Ta) to perform the task: Tr/Ta (e.g., Holley & Parks, 1987). A ratio of greater than 1 implies that the task cannot be done in the time allotted; a ratio less than 1 indicates acceptable times. This is a performance definition of workload; the task can be done within the time frame or it cannot. The Tr/Ta ratio defines an important but limited condition for overall workload definition. Normally, the application of this definition assumes an acceptable quality of performance when the task is completed, but the definition does not take into account the degree of quality of performance. Like the amount definition, inferences about workload are made from performance. If the task can be accomplished within the time available, then the operator may have spare time and spare capacity. The Tr/Ta ratio is also called time stress by some authors.

A quite different approach is to consider the time already expended. Although you will not read much about it in this volume, this is related to the effects of fatigue and the issue of workload duration. That is, a greater effort may be needed to perform an act if the person's current capacity for that action has been depleted or is currently low. Mental effort may not require great amounts of physical energy and the laws may differ for mental and physical fatigue. Nevertheless, probably everyone has had the experience of being pushed to the point that it is relatively difficult to think, leading to slower processing. Indeed, performance on a variety of cognitive tasks declined in a sustained command and control environment (Angus & Heisgrave, 1983). Mean time to process messages increased, showing the operators were working more slowly. Similarly, the number of correct responses decreased on a logical reasoning task and other tasks. However, errors did not necessarily increase on these tasks, indicating slower but equally accurate performance.

Composite Conceptualizations

Having understood the limitations with the definitional approaches described above, several researchers have suggested that workload is really a composite of several different things. For example, Jahns (1973) proposed that workload can be thought of as containing the components of

- input load,
- operator effort, and
- performance.

According to Jahns, input load is the task requirements (situation demands in Figure 2-2) imposed on the operator, that is, what is required of the operator. The second component is the degree of effort being expended by the operator to accomplish the requirements. The third component relates to operator performance and to what degree the required tasks have been accomplished.

Many other investigators also consider workload to be a multidimensional concept. Workload has been expressed as a global concept that affects operators in relation to their ability to accomplish a task (e.g., Hart, 1986b). Edholm and Weiner (as cited in Rohmert, 1987) suggest that workload is the total of all determinable influences on the working person. Therefore, all elements of work including environmental, social, motivational and other factors will affect the workload. There can be little doubt that there are individual preferences regarding what workload means and the factors that may cause it. Certainly this was the case when, for example, Hart, Childress and Hauser (1982) asked 117 people which of 19 possible components were a primary component of, were related to, or were unrelated to workload. Each of the 19 components were considered as primary by at least 25% of the individuals. However, only task difficulty and time pressure were considered a primary component by more than 72% of the raters.

Subjective vs. Objective

Amount of work and time to do the work are two objective ways of inferring workload. Somehow, however, they do not capture all there is to workload. In the driving example, performance remained acceptable throughout, but the perceived difficulty of the task increased in both time pressure and the amount of work. One would like to capture the level of perceived difficulty as an indicator of when the task will become too difficult. Workload researchers have recognized this omission and defined workload in the context of subjective and psychological variables.

Effort Needed to Perform a Task. Closely associated with the performance of a task is the effort needed to do a task. From this standpoint, workload depends not only on the particular task to be accomplished, but also the current capacity of the operator to perform the task. That is, a greater effort will be needed to perform an act, not only if the person's capability to perform that task is inherently limited, but also when the resources needed to perform the task have been partially depleted. For example, one might measure the actual physical work being done by a person doing pushups by determining the actual distances and weight being lifted. Some persons who are in better physical condition will have little difficulty in doing a certain number of pushups. For others, the same task can only be done with great difficulty. However, because of the progressive depletion of resources during this task, the final pushup may be perceived as having required considerable more effort than the first.

The concept of workload as effort also considers workload to be something internal to the operator. This makes the definition dependent not only on the normal capabilities of the individual, but also on the current states of the operator. Thus, if workload is defined as operator effort should one be measuring efforts expended, efforts anticipated, or effort currently being put forth? All three ways are implicit in formal models of the human operator, but have not always been included in definitions of workload.

Subjective Experiences When Performing a Task. Some researchers have viewed workload as subjective experience. Johanssen, Moray, Pew, Rasmussen, Sanders and Wickens (1979) concluded that, "If the person *feels* loaded and effortful, he *is* loaded and effortful whatever the behavioral and performance measures show" (p. 105). Similarly, Sheridan (1980) suggested that "mental workload should be defined as a person's private subjective experience of his or her own cognitive effort. (p. 1)."

Sheridan and Simpson (1979) have suggested that there are three categories of words that are used when talking about workload. There are words associated with task time constraints, such as the time available to complete work, the number of interruptions and the nearness of deadlines. There are also those words that are related to the uncertainty and complexity associated with a task. These include such things as uncertainty as to what the tasks are and what the consequences of various tasks will be, as well as the type and amount of planning that must be done to accomplish the task. The third kind of words are those related to psychological stress such as risk, frustration, confusion, and anxiety.

This three-dimensional definition based on time constraints, task complexity, and psychological stress was adapted and operationalized by Reid, Shingledecker and Eggemeier (1981) for use in their Subjective Workload Assessment Technique (SWAT). Time load refers to the relative amount of time available to the operator (AAMRL, 1987) and the percentage of time an operator is busy (Eggemeier, McGhee & Reid, 1983), and includes elements such as overlap of tasks and task interruption. Mental effort (task complexity) refers to the amount of attention or concentration directed toward the task, independent of time considerations. Psychological stress is the degree to which confusion, frustration, and/or anxiety is present and adds to the subjective workload of the operator. Factors that may increase stress and elevate distraction from the task include personal factors such as motivation, fear or fatigue, and environmental factors such as temperature, noise, or vibration (AAMRL, 1987).

Summary Comments

Stating that operator workload is a multidimensional concept may appear reasonable, at first glance, but it tends to beg the question of what workload really is. Workload is often used as a practical, atheoretical term. Sometimes, workload is defined in terms of the amount and number of tasks to do and the time available to do them. Instead of attempting to define the concept, these approaches tend to imply how workload should be measured and assessed. In many cases, there is little in the definition to distinguish between workload and performance. Some definitions are more internal and include psychological dimensions such as stress, effort, and difficulty. However, one can appreciate the complexity of the operator workload concept by noting all the facets that have been ascribed to it from the various definitions or conceptualizations.

What We Mean by Workload: An Analogy

Earlier in this chapter, it was pointed out that workload is not the same as performance although workload is related to performance and assumed to be a determiner of the quality of performance. It was also pointed out that there are a variety of influences on performance in the context of a human model: Performance is the coin of the realm. The various definitions of workload hint at what is deemed to be important, specifically, number of things to do, the time to do them in, and psychological factors. These points all describe performance and workload from a relatively static standpoint.

However, an operator is highly adaptable and dynamic. By putting forth more effort for short periods of time, adequate performance can often be maintained even on tasks that are too difficult or too complex to handle for extended periods of time. But high workload conditions take their toll, they deplete resources needed for various capabilities, and they may well result in inadequate performance in the future. An analogy would be if a design engineer evaluated the performance of a new vehicle only by the distance it was capable of traveling without ever considering the size of the fuel tank or the rate at which fuel was being used. The fact that the vehicle can reach long distances under some conditions does not mean that it can always reach those distances. To take the analogy further, it is also true that the relationship between engine load in revolutions per minute (RPM) and fuel consumption is non-linear. Requiring the vehicle to travel a specified distance at a very low or high RPM will use more resources per unit distance than if the same distance were traveled at an optimally efficient RPM.

These characteristics are depicted in schematic form in Figure 2-3. The ordinates show the load on the engine in RPM and distance or vehicle range as performance. In addition, the capacity of the fuel tank or amount of fuel available is represented as a parameter with several different capacities shown as curved lines. To determine the distance that can be traveled (performance), one needs to know RPM and the size of the tank. There are boundaries. RPM cannot exceed some practical maximum, i.e., the red line, if engine damage is to be avoided, and obviously, if RPM is zero, no distance will be traveled. Similarly, there are limitations of the capacity of the fuel tank; it cannot be zero and there is a practical maximum. If one wanted information about a 12.5 gallon tank, one would interpolate between 10 and 15. Performance of the vehicle under varying conditions can thus be described in non-linear terms with respect to RPM and in terms of a performance envelope which is represented as the white space in the figure.

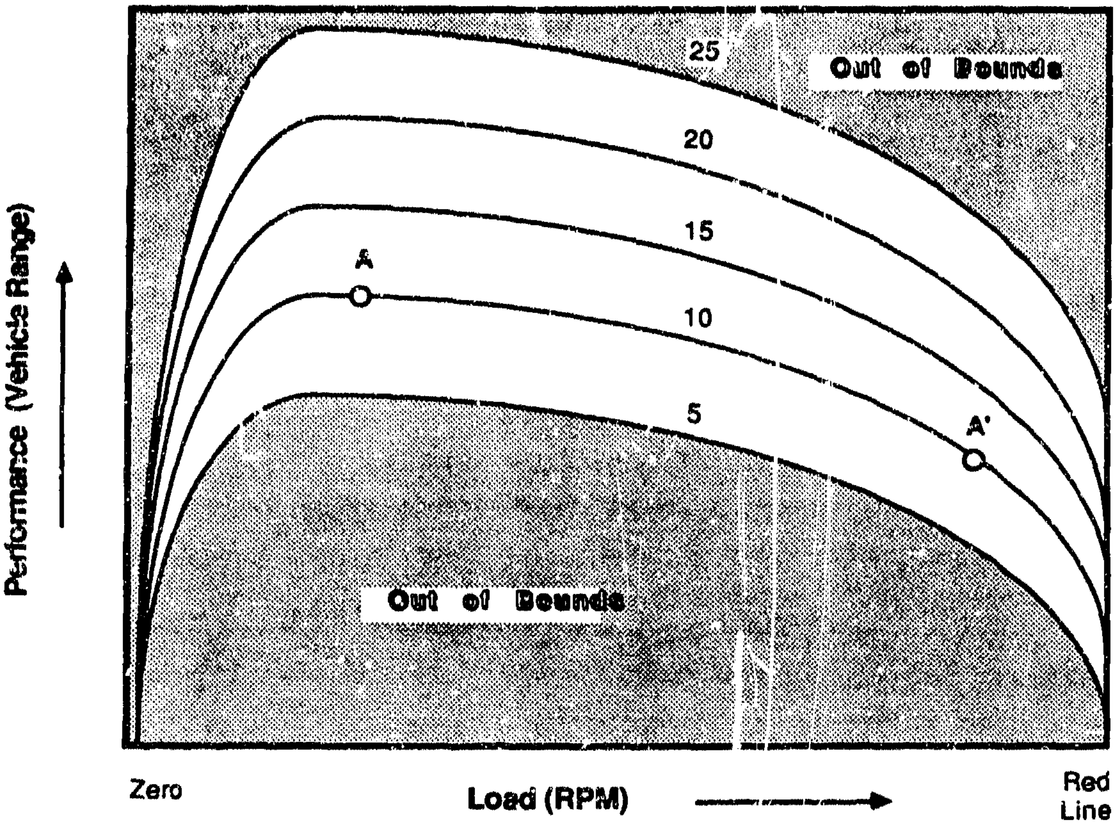


Figure 2-3. A schematic representation of a vehicle performance envelope as a function of workload/RPM with fuel tank capacity as a parameter.

There are several other points one can make in this context. Dynamic changes made in the course of execution can be represented in the figure. At or above the optimal RPM, an increase in RPM will result in a reduction of travel range which would be represented as a shift of relative position within tho envelope; for example, an increase in RPM from point A in the figure to A' results in a lower vehicle range. Similarly, one would want to plan a safety margin. For example, in aviation, the pilot is responsible for calculating the amount of fuel needed to reach the destination, plus the amount needed to reach an alternate airport, plus a further safety margin of at least 10%. In aviation, it is standard practice to stay away from the performance envelope boundaries.

In a similar way, we can consider human performance in terms of a performance envelope. We show this in Figure 2-4 which is basically a human analogy to Figure 2-3. In this case we have depicted workload (time or amount) and performanco on the ordinates. The parametor can be viewed as an estimate of the operator's current states, in short, his current capability. There is a parallel between a dynamic change in the vehicle analogy and a dynamic change in human work. Both performance functions are non-linear;

unlike the vehicle example, however, the amount of currently available human capacity *can* vary with changes in effort expended, at least up to a limit. As in the vehicle example, one would want a safety margin and that is attained by avoiding the performance envelope boundaries. But, in order to avoid the boundaries, one needs to know where in the space the operator currently is, hcw much additional work is coming in, and the rate at which this additional work will cause the operator to move toward a boundary. Thus, workload cannot be evaluated merely by knowing the amcunt of work that a task requires of the human. One also needs to know the rate at which the work must be done and the extent to which it will deplete the human resources that are available, not only for the current task, but for others that will be occurring in the future. In short, one needs to know where the operator is in the performance envelope at any given time.

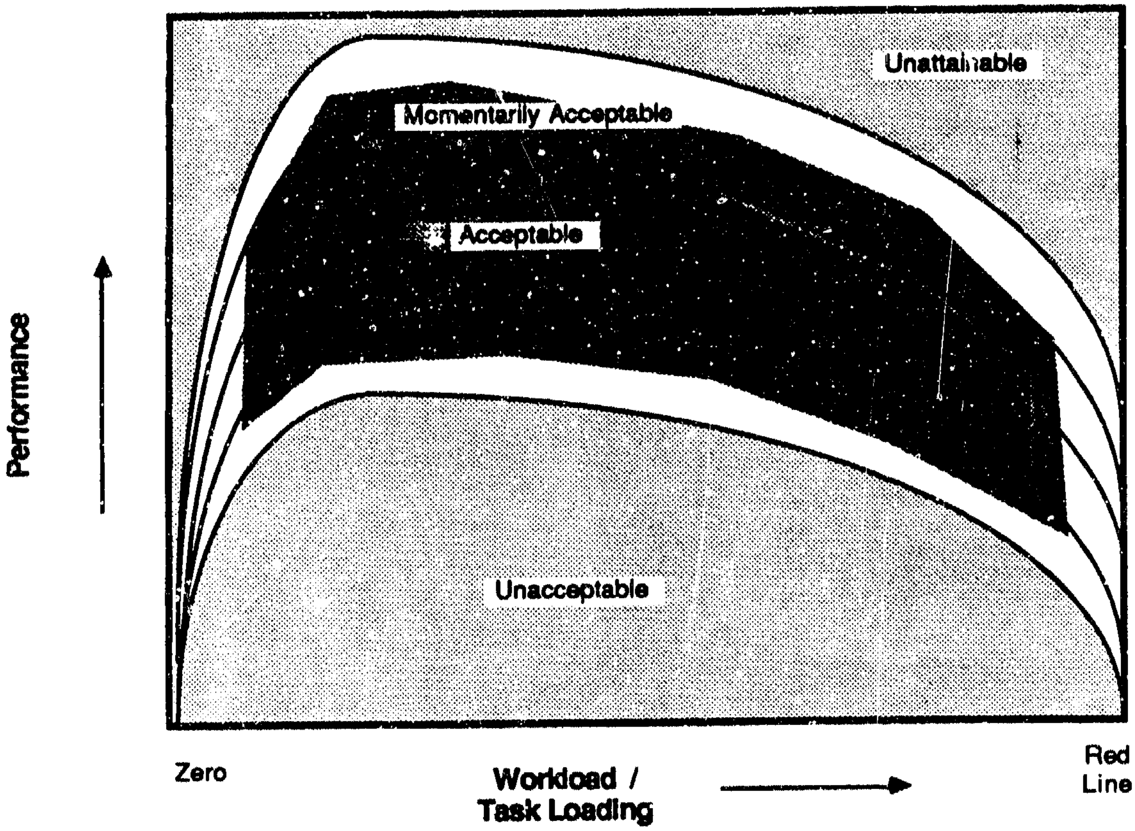


Figure 2-4. A schematic representation of human performance and the workload envelope.

Nor is it sufficient merely to consider the impact of various tasks on the average operator. Individuals differ in their capabilities and resources at the beginning of a mission, and those differences may become more pronounced as the mission unfolds. Operators who start a mission with lessor capabilities may have to expend their limited resources faster than those who began with greater capabilities. Task demands and the likelihood of humans being able to accomplish them cannot be analyzed and evaluated without considering both the individual and the impact that previous tasks may have had on the individual. Just as we are interested in the readiness of a military unit to respond to various types of demands that may be put on it, the practitioner should be interested in the moment-to-moment readiness of the individuals to respond to various task demands. That is, we need to know the starting position (the capacities of the mental and physical fuel tanks) of the operator in the workload space and how close the operator is to the boundaries of the envelope.

What We Mean by Workload: Describing the Elephant

The review of the definitions did not indicate any overwhelming unanimity among the authors. Indeed, the definitions reviewed have more in common with assessment technique descriptions than with conceptual definitions *per se.* Having made these few points, let us be venturesome and extract some conceptual principles concerning workload. Our tenets of workload are:

- Workload is relative. It depends on both the external demands and the internal capabilities of the individual. This relativity exists in both dimensions of amount and time, e.g., it can vary over time for an individual.
- Workload causes the individual to react in various ways. Workload is not the same as the individual's performance in the face of work or tasks.
- Workload involves the depletion of internal resources to accomplish the work. The higher the workload, the faster resources are depleted.
- There are a diversity of task demands and a corresponding diversity of internal capabilities and capacities to handle these demands. Persons differ in the amount of these capabilities that they possess.

Out of these tenets we can derive a working definition of workload. It is not the intention here to propose *the* definitive meaning, but rather to suggest the working definition for the purposes of understanding and of practical application. In the sense that workload and performance are related in the manner shown in Figure 2-1, what is really of interest is to predict that point just short of rapid degradation of performance. This can also be stated in terms of the current vs future position of the operator in the workload envelope of our analogy in Figure 2-4. Past performance can be measured, but the future ability of the operator to perform is what the practitioner would like to know. Where in that hypothetical performance envelope does the operator currently lie? In this sense, the aspect of workload that needs most to be estimated or measured is considered to be *the relative capacity to respond.* This working definition is meant to imply not only the amount of spare capacity, but also the ability of the operator to use that capacity in the context of the specific personal and environmental situation.

By proposing a working definition as *the relative capacity to respond*, the emphasis is on predicting what the operator will be able to accomplish in the future. It is a global definition in that it does not necessarily attempt to explicate the specific factors or dimensions that will influence individuals in their performance or perception of workload. (The definition is, however, consistent with all the points made.) At all times, workload will involve the interaction of the operator with the task and these two elements cannot be separated totally. At the same time, the circumstance will dictate to what extent operator characteristics or task characteristics will be important in the assessment of workload. The specific situation will determine the most appropriate questions to ask about operator workload, and consequently the most appropriate ways to answer those questions.

Taxonomies of Workload

Our working definition cuts across the various techniques used in workload assessment. To discuss the techniques we need a different framework, and for that organizational framework, we utilize a taxonomy. Taxonomies are developed as aids in classification. Classification serves the useful purpose of grouping similar items together as well as being helpful in explicating their structure. Researchers have used various workload taxonomies for the two main purposes of (a) classifying the nature of the operator tasks and (b) classifying workload assessment techniques.

Task taxonomies are useful because some workload techniques appear to be able to discriminate high and low levels of workload in some types of tasks better than others. Often this differential discrimination results from the specific design of and the intention behind the technique. A task taxonomy can be useful in helping to determine the more appropriate workload techniques for a specific application.

Taxonomies also have been developed to classify workload methods and techniques into descriptive categories. As will be discussed later, some categories of methods are more useful for specific circumstances than others and the classification scheme provides a convenient vehicle for categorization. By classifying both tasks and techniques, matches may be found more easily.

By Task. A comprehensive review of the operator workload literature was completed nearly a decade ago by Wierwille and Williges (1978). In the report, they provided a survey and analysis of 400 workload studies. For classification, they used a human operator task taxonomy (called Universal Operator Behaviors) that had been developed earlier by Berliner, Angell, and Shearer (1964). In this task taxonomy, human activities in systems are separated into four broad categories:

- Perceptual tasks or sensing tasks; for example, seeing a warning light on an instrument panel;

- Mediational or cognitive tasks are those that involve thinking (e.g., solving mathematical problems);

- Communication includes face-to-face speaking, radio, and other communication tasks; and

- Motor processes are those which involve muscles or body movement (e.g., activating a pushbutton).

The Universal Operator Behaviors taxonomy has been adopted by other workload researchers as a useful task taxonomy. (A complete description of the taxonomy appears in Chapter 8). Some of these categories will be discussed indirectly in the context of the review of techniques.

By Technique. Wierwille and Williges (1978) separated workload techniques and measures into four categories, namely subjective opinion, spare mental capacity, primary task, and physiological measures. Having developed these categories, they used them to categorize workload techniques with respect to operator behaviors. Other taxonomies of workload have been developed as well, including subjective/objective subjective/performance/physiological, and similar variants. For example, Johanssen (1979) suggests a four-group classification for techniques to measure operator effort: time-line analyses, information processing studies, operator activation-level studies, and subjective effort ratings. Moray (1979a) suggests that OWL techniques be divided into normative, physiological, and empirical measures corresponding to the three components of the structure suggested by Jahns (1973), specifically, input load, operator effort, and performance. As suggested by Moray (1979a), normative measures include those which look at the input load, such as queueing theory; physiological measures include those that attempt to measure the effort or activation level involved, such as heart rate or EEG; and, empirical (behavioral) measures are those related to performance such as reaction time or root mean squared [RMS] error.

Other researchers suggest classification schemes with more categories. For example, Strasser (Hamilton, Mulder, Strasser, & Ursin, 1979) has developed a taxonomy of OWL methodologies with eight categories:

- Vegetative variables – heart rate, blood pressure, respiration, galvanic skin response;

- Central nervous variables – electroencephalogram, evoked potentials;

- Biochemical variables – hormone levels in bodily fluids;

- Peripheral variables – pupil diameter, electrooculogram, critical flicker fusion frequency;

- Subjective methods – rating scales;

- Loading tasks – continuous and discrete; paced and self-paced tasks;

- Performance measures – reaction time, etc.; and

- Observations – task analysis and behavioral measures.

The categorization of techniques in this taxonomy reflects a particular interest in physiological measures of workload (Strasser, 1937). Clearly, classification schemes are created to meet specific needs of the researcher and the intended user and application.

An Expanded Technique Taxonomy

The workload technique taxonomy used for this report is shown in Table 2-1. It is designed to be flexible and meaningful in addressing OWL issues in the Army. It differs from previous taxonomies in that greater emphasis is placed on those analytical techniques that can be used to predict OWL during system concept development and preliminary system design. Previous taxonomies and most OWL research have concentrated on empirical techniques that are applicable only at more advanced stages in the system development cycle - - i.e., they are test and evaluation oriented rather than design oriented. The term analytical is used to label techniques which are used in a predictive manner without actually employing an operator; operator-in-the-loop techniques are labeled empirical. It is quite clear that the Army needs both types of techniques. The taxonomy is elaborated in the discussion of classes of techniques in subsequent chapters and presented in detail to include all techniques in Chapter 8.

Analytical Techniques. The focus of the analytical techniques is on workload analysis that may be applied without operators-in-the-loop. These techniques are used to predict workload early in system development where the greatest design flexibility is available with the least impact on system cost. These techniques may also be used throughout hardware development to guide, augment, or extrapolate beyond operator-in-the-loop investigations. The analytical techniques are classified into five categories: (a) Comparison; (b) Expert Opinion; (c) Mathematical Models; (d) Task Analysis Methods; and (e) Simulation Models. These analytical categories are discussed in detail in Chapter 3.

Empirical Techniques. The empirical techniques have received considerable attention and are the most familiar methods (O'Donnell & Eggemeier, 1986). The taxonomy of empirical techniques presented here includes four major categories (O'Donnell and Eggemeier, 1986) and is similar to that developed by Wierwille and Williges (1978) . These include: (a) Primary task measurements which focus on the degree to which human and system performance achieve stated goals. (b) Subjective methods that assess operator opinion and include rating scales as well as questionnaires and interviews. (c) Secondary task approaches have been used to examine the amount of operator spare capacity. (d) Physiological techniques, both classical (e.g., heart rate) and specialized (e.g., heart rate variability or evoked potentials) which continue to be examined as to their most appropriate application in workload assessment. These classes of techniques are discussed in Chapters 4, 5, 6, and 7, respectively.

Table 2-1. Taxonomy of workload assessment techniques.

TECHNIQUE	CATEGORY	SUBCATEGORY
Analytic	Comparison	
	Expert Opinion	
	Math Models	Manual Control Models Information Theory Models Queueing Theory Models
	Task Analysis Methods	
	Simulation Models	
Empirical	Primary Task	System Response Operator Response
	Subjective Methods	Rating Scales Questionnaire/Interview
	Secondary Task	Subsidiary Task Probe Task Dual Task
	Physiological	Classical Specialized

Some Additional Definitional Issues in OWL

There are some additional definitions and conceptual tools useful for workload analysis. Because of their relevance for analysis and workload assessment, they are addressed in this section. The issues tend to be more important for empirical techniques than for analytical techniques; however, they are relevant for both. Analytical techniques use definitions to identify performance and workload measurement. The developer and user can decide in a relatively direct manner how he wants to assess workload. With empirical techniques, however, the issue is not quite as straightforward. It is not easy to go back to collect data that were missed on the first test. Nor is it easy to scrap a technique and replace it with another that provides a more desirable level of quality and detail. Clarification of these concepts at this point will help the reader in evaluating the subsequent review and discussion.

Sensitivity of Techniques and Measures

Sensitivity of workload assessment techniques is the degree to which the various techniques can differentiate between levels of load placed upon the operator. Some investigators (e.g., Wierwille et al., 1985) have stressed issues of workload assessment sensitivity. It is generally accepted, mistakenly, that most empirical workload estimation techniques are sensitive to changes in load imposed on or experienced by an operator. In fact, the majority of techniques are insensitive when tested in scientific experiments. For example, Wierwille and his colleagues tested 25 different techniques in four experiments and found that only about 25 to 30% of the techniques had any usable sensitivity. However, the sensitivity also depends on the appropriateness of the technique for the system.

Lack of sensitivity is the single most critical issue in selection of an empirical technique. If an insensitive technique is used, it will indicate there are no changes in workload regardless of the values of the independent variables. This could lead to systems with workload problems discovered only after fielding. It is for this reason that we advocate using multiple techniques when assessing workload.

Diagnosticity

Diagnosticity refers to the extent to which a technique reveals not only overall assessment of OWL but also information about component factors of that assessment. For example, an important diagnostic is the ability of a measure to differentiate among various sensory, perceptual, cognitive and psychomotor aspects of human performance. The concept as used in workload has been attributed to resource theory (O'Donnell and Eggemeier, 1985) but the basic methodology for such a differentiation can be traced back to Garner, Hake and Eriksen (1956). The essence of the notion of diagnosticity is to be able to identify the specific mechanism or process involved or overloaded during performance of a particular task. Typically, the diagnosis is an inference based on the information available. Garner et al. (1956) have formalized the concept of converging operations, a diagnostic methodology for attacking the problem in several different ways to insure the quality of the inference. The converging operations method is critical to the diagnosticity of workload techniques.

Diagnosticity is an important issue, but most workload measures are inherently weak in this regard. Diagnosticity can often be improved significantly by simultaneously recording system changes induced by the operator, i.e., recording control inputs or other observable behaviors. This is a version of converging operations. An example is provided by Harris and Christhilf (1980) in which control inputs of the operator were recorded simultaneously with eye movements. This allowed the investigators to relate control inputs to dwell times. They found that longer fixation time or dwell time on an instrument were associated with control inputs while shorter times on the same instrument were not associated with control inputs. Thus, a long dwell time without a control input would imply difficulty in interpreting the instrument and in turn would implicate a cognitive mechanism. By inference, more mental activity and therefore more decision processes were associated with the longer dwell times. Other analyses are consistent with this suggestion (Dick, 1980). One measure by itself would not have permitted such an inference.

Technique vs. Measure

A *technique* is a generic term referring to a workload assessment methodology. A *measure* is a specific assessment scale or a metric. For example, collecting heart data with either a wrist band for pulse or chest electrodes qualifies as a technique. Scoring the data for mean heart rate qualifies as a measure; it is a form of a metric and data analysis applied to heart data. (It may also involve considerably different assessment scales.) Similarly, evaluating heart rate variability is another measure or metric applied to data collected with a heart technique.

When selecting empirical estimates of performance to derive workload, an investigator must not only choose appropriate techniques, but also appropriate measures. Within a technique, sensitivity may vary with the measure selected. For example, if time estimation has been selected as a technique to be used, there are many measures that could be employed: absolute error, standard deviation of estimates, root mean squared (RMS) error, or number of no-response intervals. Technique sensitivity is often dependent upon the measure used. Wierwille and Connor (1983) and Savage, Wierwille, and Cordes (1978) demonstrated this for two secondary task techniques (i.e., time estimation and digit shadowing). Measures should be selected carefully and should be based upon previous research or preliminary investigation.

Technique vs. Procedure

A *procedure* is the application of a technique specifying the steps taken in applying a technique. This is like plotting out two different routes to get from point A to point B. Differences will be in terms of the quality of the ride, the time taken, the likelihood of getting lost, etc. Similarly, a given technique can be applied in different ways and each variation may affect performance differently. For example, time estimation can be used in many ways. The following are examples of procedural variations:

- Subject produces intervals x seconds long, or x seconds after an auditory or visual signal.
- Instructions indicate whether task is to be neglected under high load or to be performed regardless of load.
- Interval produced is to be 5, 10, 15, 20, or 25 seconds.
- Subject's response is verbal, pushbutton, or pedal actuation.
- Subject is instructed to count or subject is instructed not to count.

Because performance, as well as sensitivity and diagnosticity of the technique, is affected by procedure as well as by technique and measure, each aspect of a procedure should be considered and decided upon before actual data collection. Procedural aspects should be based on results reported in the literature and application specifics. As with the second example at the beginning of this chapter, a procedural change as simple as altering the mode of response from verbal to written responses can make a big difference in both performance and the subjective experience of the respondent.

Overview of Analytical Techniques

An analytical technique produces results that are used to predict performance and estimate workload without actually having a human operator exercise the system. This definition of analytical technique applies even when a potential operator of the system under development, or the operator of a similar system, may offer expert opinion as a subject matter expert (SME). In contrast to analytical techniques, empirical techniques are those which require a human operator to interact with the system in question. The identification and development of useful analytical procedures for estimating workload and predicting performance continues to be an actively pursued goal. This is especially true in the applied sector, where system developers need to assess workload early in the design process while conceptual designs are easily modified.

The general difficulties that exist with the assessment of OWL (as described in Chapter 2) are most pronounced for analytical techniques. The lack of operator interaction with the system presents problems in defining the relevant workload issues and measures. There is also the added difficulty of the scarcity of detailed data about the system that is to be operated by the human. Typically, analytical techniques predict performance and potential performance failures. Workload, therefore, is often an inference derived from a prediction that a task cannot be performed to criteria or standards. For example, an operator's activities may require more time than is available within the time constraints and requirements of the mission.

There is no fully accepted formal model defining the factors which drive workload nor relating the contribution of each factor to overall workload and its subsequent impact on performance. The result of this deficiency is that various analytical techniques use different measures to assess workload. Some techniques estimate workload without explicitly considering the human, defining workload in terms of task demands such as numbers of tasks to be performed. Others try to estimate the operator's attentional reserve capacity, following theoretical constructs of human abilities. Finally, some analytical techniques attempt to incorporate empirical or observed human performance capabilities within the workload estimation process.

Few, if any, of the available analytical approaches may be considered to capture the full complexity of the workload issue. However, the techniques cover a variety of workload issues. Thus, each individual method can provide the developer some useful OWL information as well as information about the operator and system performance. In general, two conclusions may be drawn about analytical techniques in particular, and OWL techniques in general:

- A battery of techniques, both analytical and, if possible, empirical, is needed for each situation.
- Different situations require a different mix of OWL assessment techniques.

A number of very useful techniques have evolved. Some are more general than others, and some more applicable to certain problem domains than to others; the difficulty is to determine which techniques are best suited for a specific application. The intent of this chapter is to describe the various analytical procedures, assess the utility of each, and provide specific examples of each procedure. Table 3-1 comprises five major categories of workload estimation techniques, each of which is described in detail in subsequent sections of this chapter. The first class of techniques involves comparison with predecessor or reference systems. The second technique, expert opinion, involves the elicitation of workload estimates and predictions from operators or other system experts. Third, mathematical models represent attempts to abstract and quantify aspects of the human-machine system through the use of formal mathematical representations and relationships. Fourth, task analysis techniques, based on detailed decompositions of the intended missions into individual tasks, are described. Lastly, approaches to computer simulation of human performance are considered.

Table 3-1. Taxonomy of analytical techniques.

ANALYTICAL TAXONOMY

- Comparison
- Expert Opinion
- Mathematical Models
- Task Analysis Methods
- Simulation Models

A Summary Evaluation of Analytical Techniques

In addition to a review of the techniques, there is an intent to provide guidance on which procedures may be best suited to a given set of resources and measurement goals. Toward that end, Table 3-2 provides an overview of the techniques and a consensual judgment of the present authors about the data

Table 3-2. Comparative overview of the analytical techniques.

Technique	Data Requirements	Cost/Effort* Requirements	Diagnosticity	Subjectivity
Comparison	System level	Low cost/ Low effort	Low	High
Expert Opinion	Task level Low effort	Low cost/	Low-Moderate	High
Math Models	Task level High effort	Low cost/	Low-Moderate	Low
Task Analysis				
Time Based	Task level	Low cost/ Moderate effort	Low-Moderate	Low
McCracken-Aldrich	Task level	Low cost/ Moderate effort	Low-Moderate	Moderate
Simulation				
Siegel-Wolf	Task level	Moderate cost/ High effort	Low	Moderate
SAINT	Task level	Moderate cost/ High effort	Low-Moderate	Moderate
Micro SAINT	Task level	Low cost/ Moderate effort	Low-Moderate	Moderate
SIMWAM	Task level	Moderate cost/ Moderate effort	Low	Moderate
SWAS level	Task element Moderate effort	High cost/	Low	Moderate
HOS	Task element level	Low cost/ High Effort	Moderate-High	Low

* Cost refers to acquisition costs in dollars. Effort includes number of personnel and development time/effort.

requirements, costs, diagnosticity, and subjectivity of each technique.The column entries are defined as follows. The term data requirements refers to the level of detail required to use the technique. These range from system level data for comparison down to the task element level for simulations. Cost refers to the acquisition cost of the technique while effort refers to the relative number of human hours needed to apply the technique. Diagnosticity gives an estimate of how well the technique will pinpoint causes of workload. Subjectivity refers to the amount of judgment required on the part of the user and/or SMEs. The potential user may consult this table as a guide to identify techniques of particular interest, and then pursue additional reading for more information.

Comparison with Existing Systems

New system development is traditionally more evolutionary than revolutionary. Typically, an enemy's technological developments or increased level of threat requires upgrading or replacing older weapons systems with newer versions that perform essentially the same functions. In this case, the older system can provide an abundance of lessons learned, if that information can be obtained in a useful format. The comparison method uses the physical and functional similarities between existing and proposed systems to extrapolate data from the fielded system and apply them to the conceptual system. There is little published material describing the application of comparison to workload issues, although some techniques have been developed in allied areas. However, more formal techniques for this comparison process must be developed if its full potential is to be realized. Relevant work which has been reported is briefly summarized below.

Use of Comparison for Predicting Workload

A systematic attempt to use a comparative technique for predicting OWL is that of Shaffer, Shafer and Kutch (1986). They developed workload estimates for a single crew light experimental helicopter (LHX) scout mission. They based their estimates on an earlier detailed, time-based workload analysis of scout missions conducted in a OH-58D helicopter with a two-person crew. Had a good workload database already existed on the OH-58D, their comparison might have been performed more easily and effectively. Nevertheless, their effort represents one of the first attempts to systematically compare conceptual and existing systems in terms of OWL.

Use of Comparison for Predicting System Effectiveness

John, Klein and Taylor (1986) have developed a formalized comparison method for evaluating a system by using analogical reasoning based upon what is known about a comparable system. Their method, known as Comparison-Based Prediction (CBP), is an extension of Comparability Analysis used by the Air Force to estimate system reliability and logistic requirements. CBP is essentially a technique for structuring and quantifying SME opinion and involves identifying factors that are expected to influence relevant system characteristics of interest. Comparison cases or systems are then selected and rated as to whether they possess more or less of these characteristics. The causes of these judged differences are then examined ultimately, to identify adjustment factors that can be applied to the comparison system operational data to produce predictions for the system under study. In cases where applicable operational data do not exist, they can be generated by SME estimates, although this will reduce confidence in the results obtained.

CBP feasibility studies were conducted to develop estimates of the training effectiveness of three training devices: automotive maintenance task trainers, tank gunnery simulators, and howitzer trainers (John et al., 1986). These studies indicated that CBP was a viable estimation technique that was useful in generating design recommendations. While CBP has not yet been applied to workload explicitly, the authors state that it could "... enhance a preliminary subjective workload assessment model by providing reference anchors in comparable equipment, existing metrics, and operational experience" (p. 152).

Early Comparability Analysis in Manpower, Personnel, and Training (MPT)

The Army MANPRINT initiative encourages the use of predecessor or reference systems in the analysis of anticipated new system requirements (U. S. Army, 1987). To that end, an Early Comparability Analysis (ECA) methodology has been developed (U. S. Army Soldier Support Center, 1986) to identify MPT requirements early in the material acquisition process. A baseline comparison system, either an actual whole system or a composite system made up of applicable components of other systems, is defined and used to establish high driver tasks. These tasks which significantly impact MPT concerns help to define the expected number and types of people or the required amount of training. The MANPRINT initiative may be expected to promote the use of comparability assessments of OWL.

Summary

The advantage of the comparison technique for predicting OWL is its ability to obtain more rigorous data than purely subjective estimates. Currently, comparison is less of a well defined technique than it is a generalized procedure. Although one might like to see empirical workload data used as the basis for estimating the workload on the conceptual system, such a data base could also be obtained from validated analytical techniques such as task analysis. These data can be existing, or collected specifically for comparison purposes. Unfortunately, most current operational systems do not have a workload database, and for those systems that do, the data often have questionable reliability and validity.

Thus, the comparison technique offers a fairly straightforward analysis, but only if data are available on a predecessor system. The *if* seems to loom large. While it is likely that the technique is often used informally (and overlaps the expert opinion technique), there appears to be a lack of documented applications. One major impediment to making comparison analysis a viable technique is the lack of systematic databases on existing systems. However, as operator-in-the-loop workload evaluations of existing systems become more of an established practice, use of comparative techniques to estimate new, derivative system workload should be facilitated. For example, a good, solid database is being built for helicopter evaluations (e.g., Szabo, Bierbaum, & Hocutt, 1987) which will make OWL comparison much easier for helicopters. If similar databases are constructed for other types of systems, the comparison techniques may be expected to have growing utility.

Expert Opinion

Expert opinion is the oldest and most extensively employed workload prediction technique. This is probably due to several factors including ease of implementation, relatively low cost, and a large supply of experts. The first part of this approach, given a system defined to some preliminary level of detail, is to identify the users or developers of systems that are either predecessors or functionally similar to the system under study. These individuals or subject matter experts are then given a description of the new system and its intended use, perhaps within the context of a detailed operational scenario. The next step is the elicitation of the subjective opinions of the SMEs on how the system might perform, focusing on major strengths and weaknesses. Analytical evaluations of workload may be developed through this approach in a manner similar to that used in the comparison method. Employed as described, this technique provides a capability to identify broad workload problem areas early in the design process.

The application of the expert opinion technique described above is usually relatively informal. Often, it is of considerable benefit for the workload analyst to have an expert describe the details of operation in an unstructured manner. However, this informality may introduce considerable variability into the quality of the results obtained. Whether due to levels of experience, familiarity with types of systems, verbal capabilities or SME bias, individual differences in SMEs can produce a substantial spread in the workload estimates. There also may be miscommunication between the investigator describing the system and the SMEs, resulting in an erroneous understanding of how the system operates. Fischoff (1983) provides a good overview of the problem of eliciting expert opinion. For this technique to be more objective, a structured, formal approach is needed in both the selection of SMEs and the elicitation of information.

Delphi Technique

Attempts to structure expert opinion have been made; the Delphi method, for example, has been developed for reducing the variability in SMEs' workload estimates (Dalkey, 1969). This technique is "...a process whereby subjective judgements or the implicit decision-making processes of experts can be made more objective and explicit" (Meister, 1985, p. 423). Generally, Delphi is administered to a group of SMEs. The eventual goal is to arrive at a group consensus, for example, on the expected workload for the defined system and scenario. The Delphi Technique involves several phases, most of which are iterations or rounds in which the results of previous rounds are summarized and returned with a questionnaire to the group of SMEs. The method is most applicable to situations in which existing referents or comparison systems are not available, or where extrapolation or prediction are required. The validity and reliability of the Delphi method is subject to the same constraints as any other subjective method, but where such methods are required, the more structured Delphi method may strengthen the results.

Prospective Subjective Techniques

The most significant systematic effort in expert opinion has been the development of an analytical, prospective application of the Subjective Workload Assessment Technique (SWAT), dubbed Pro-SWAT (Reid, Shingledecker, & Eggemeier, 1984). Because less work has been done using Pro-SWAT, we defer discussion of most of the details of its development and application to Chapter 5 which describes SWAT. Like SWAT, Pro-SWAT has a scale development phase and an event scoring phase. The procedural outline of a Pro-SWAT session, as described by Kuperman and Wilson (1985), involves the following steps:

- Define workload and describe SWAT and Pro-SWAT.
- Develop the measurement scale.

- Describe the mission equipment package including controls and displays (i.e., the switching logic and formats).
- Provide an overview of the mission scenario segments that comprise the role playing exercise.
- Execute role playing - Run through the scenario using whatever props are available.
- Obtain Pro-SWAT ratings - Obtain ratings after completion of each significant task or mission segment.
- Conduct a structured debriefing.

Pro-SWAT has been applied to a variety of systems. Acton and Crabtree (1985) used it to evaluate an improved version of a military C^3 system; Detro (1985) Eggleston (1984), Eggleston and Quinn (1984) all describe applications to advanced aircraft systems; and Kuperman (1985) describes the use of Pro-SWAT in evaluating advanced helicopter crewstation concepts. Eggleston (1984) compared Pro-SWAT and SWAT workload ratings provided by two separate groups of pilots. One group participated in a Pro-SWAT exercise using several configurations of an advanced attack aircraft, the other group flew these configurations under the same scenarios in a flight simulator. A Pearson correlation coefficient of .85 was obtained between Pro-SWAT and SWAT, indicating a high degree of agreement between the analytical and empirical techniques.

Summary

In summary, the utility of the expert opinion techniques for OWL prediction is high during initial stages of system design. Evidence from the studies reported above suggests their use in situations when more objective methods are not applicable, and formalizing expert opinion, as represented by the Delphi technique, helps the SME to define workload more objectively.

Theoretically, any empirical subjective assessment technique such as SWAT could be used as an analytical technique and performed prospectively. Doing so would provide a more structured process for eliciting expert opinion. However, the results would be subject to the same caveats as the parent empirical technique, as well as considerations based on the introspective nature of SME estimates.

Mathematical Models

One of the earliest goals of researchers in workload-related areas was to develop a rigorous mathematical model which would be useful for predicting operator/system performance. In principle, such a model would identify the relevant variables and combine them appropriately so that workload-associated effects on performance could be accurately and reliably estimated or predicted. The major steps, as in all attempts to model human performance, were to:

- Identify variables that influence workload either directly or indirectly.
- Determine the lawful relationships by which these variables combine.
- Establish how the resultant workload predictions drive predictions of performance.

To date, no fully comprehensive mathematical model has been developed. Several investigators have taken existing models from engineering application domains and extended them to some aspect(s) of workload-related operator performance. The most prominent of these models are based on manual control, information theory, and queuing theory. Each model is proposed to contain some parameter or component that reflects the operator's load or effort under specified conditions. Some models contain specific parameters that are proposed to be an index of load; others presume loading by defining the environmental input characteristics that are assumed to affect OWL and performance. The assumption in both cases is that these models will predict workload-related drivers and resulting performance.

Many of the models described below are aimed at continuous control tasks or information monitoring tasks which have information presented on separate displays. In part, this is because these tasks have been and still are important in complex system control. More importantly, the associated performance characteristics are definable and thus are amenable to this level of mathematical modeling. Today, with greater use of automated flight control systems and multifunction information displays, the manual control task characteristics are becoming relatively less important. This does not mean, however, that operator workload is concommitantly reduced. Indeed, the reverse is true. The implication is that mathematical models need to be developed that reflect the current set of increasingly cognitive tasks.

Manual Control Models

The manual control models fall into two general categories, those based on classical control theory and those that use modern state-space estimation methods as exemplified by the optimum control model. Both were developed within the context of continuous manual control tasks, such as piloting a vehicle. Consequently, their application to workload estimation and prediction is generally restricted to environments involving continuous controlling tasks. Designers attempt to model the human operator engaged in such a task so the combined human-machine system performance may be determined. The resultant model reflects the effort (workload) the operator is expending in order to maintain control of the system. Extended treatments of both of these types of models can be found in the literature (e.g., Kelley, 1968; Sheridar. & Ferrell, 1974; Rouse, 1980). For an excellent treatment of behavioral aspects of control theory see Pew (1974).

Manual control models have proven extremely valuable in aircraft system development where accurate prediction of handling qualities is essential to development of flyable aircraft. Although these models may be adapted to estimate measures associated with OWL in this context, the mathematical sophistication required to develop or even understand the models limits their applicability. Detailed system parameters must also be provided to exercise these models fully; these parameters are frequently not available during early concept development. Consequently, manual control models are not viable for many conceptual system evaluations.

Classical Control Theory. Classical control theory uses closed loop stability analysis methods to generate describing functions of the human operator engaged in a continuous control task. In essence, the human is considered to be a servomechanism attempting to eliminate perceived errors. Error, such as deviation from path, is the input to the model, and operator response via some manipulator device is the output. These models provide a continuous prediction of operator output over time. In workload estimation applications, a baseline operator describing function is developed. External loading factors are then applied which change the characteristics of the model in a manner which is believed to be indicative of workload. For example, system response lags to operator control inputs can be varied. Changes ascribed to increased loading may be used to predict OWL to the extent that the conditions under which the describing function was developed are generalizable.

An application of classical control theory to the workload estimation problem is described in Hollister (1986). A model is developed to estimate the allocation of an aircraft pilot's attention among continuous control and a number of other managerial tasks. The model provides insight into the nature of control task degradation due to divided attention through changes in the describing functions. It also provides an indication of the attentional demands required for control activity and the excess capacity left for managerial tasks. The stated assumption is that bad handling qualities leave little capacity for managerial tasks; good handling qualities leave more capacity. System design goals are to maximize excess control capacity. For example, to reduce the attentional demand for primary flight control, displays can be redesigned so that less time is required for gathering flight information. Despite the ability of the model to predict performance, it is generally limited to continuous control workload. However, the model has been able to predict pilot ratings of aircraft handling quality.

Optimal Control Model. Modern control theory uses sets of differential equations containing state variables and control variables to describe the controlled system. This state-space estimation theory has produced the optimal control model (OCM). An optimal controller, when given a process to control, does so by (a) observing the state variables to the degree of accuracy possible, and (b) generating a control response to these variables while minimizing a performance criterion or cost function. The criteria are usually defined as a function of error, control effort, or time. The OCM assumes that a well trained human operator will behave as an optimal controller. This implies that the operator will be aware of his own and the system dynamics. That is, the operator has knowledge of human response capability, the disturbances affecting the system, and the criterion which defines optimal control. Variables such as observation noise and motor noise are used to introduce error (Baron, 1979) and can be related to attentional scanning which is one variable considered to reflect difficulty, and hence workload. OCMs of the human operator have performed reasonably well in matching observed behavior and are capable of handling complex multivariable systems (Baron, 1979). Within the appropriate context, the predictive validity of these models makes them very useful, although their mathematical complexity makes them inaccessible to most investigators.

An excellent treatment of applications of OCM to workload estimation may be found in Levison (1979). In this report, Levison traces the development of the model, defines the basic workload model, cites a number of validation studies, and suggests issues for further development of the model. Additional examples of the model's application can be found in Rickard and Levison (1981) for the prediction of pilot ratings of the handling quality of different aircraft configurations, and in Wewerinke (1974) and Smit and Wewerinke (1978). These applications of OCM predict a workload index based on control effort which is developed in terms of OCM parameters. Levison (1970) defines an OCM model containing an attention parameter which influences the observation noise within the state variable estimator. This parameter can be used to determine the attention allocated to a display variable and hence the relative importance of that display variable in a control task. The OCM model can also be used for display design evaluation (Baron & Levison, 1977; Gainer, 1979).

A recent development of the OCM approach is the Procedure-Oriented Crew (PROCRU) Model (Baron, Zacharias, Muralidharan, & Lancraft, 1980). PROCRU provides a framework for dealing with both discrete and continuous tasks. In a discrete task application, Levison and Tanner (1971) replaced the control law with a Bayesian formulation and were able to simulate human performance for detection of a signal in noise. The OCM has considerable breadth and most of the studies have corresponding validation data. OCM is clearly a performance model with parameters which represent workload manipulations. These manipulations are of the form of amplitude, frequency, or phase lags in the equations. As a result, workload definitions are as varied as the manipulations employed.

Information Theory Models

Information theory as applied to models of human activity achieved its height of popularity during the 1960's. A good general treatment of information theory can be found in Sheridan and Ferrell (1974). Applications of information theory in psychology can be found in Attneave (1959) and Garner (1962).

Information theory provides a metric of the transmission of information through an imperfect communication channel. The metric is stated in terms of the log (base$_2$) of the number of alternatives weighted by their probabilities of occurrence. Information transmission is a reduction in the number of alternatives which is expressed as a reduction of uncertainty. Two alternatives which contain common information are said to be redundant. The channel imperfections are defined, for example, as noise and limits of channel capacity which result in lost information (equivocation).

One of the first applications of information theory to the workload domain was that of Senders (1964). In this application, a model was used to describe the division of attention by an operator monitoring information displays. It assumed that an operator, with a limited input channel capacity, sampled each information display at a frequency necessary to reconstruct the signal being presented on that display within specific error tolerances. The amount of time spent sampling each instrument is summed over all instruments to determine the fraction of the operator's time that must be spent observing. This time fraction is used as a measure of visual workload imposed by the information displays.

The use of information theory in the analysis and estimation of workload has been limited. Despite some efforts (e.g., Crawford, 1979; Rault, 1976), applications in realistically complex environments are difficult to achieve due to the necessity of *a priori* establishment of the relevant simple and conditional stimulus and response probabilities. Because information theory provides output with respect to steady-state situations, it is not well suited for representing dynamic changes in workload. The impact of information theory is probably most strongly felt through the adoption of its concepts such as limited channel capacity, information transmission, redundancy, and other concepts now contained in information processing approaches to behavior (Garner, 1974).

Queuing Theory Models

Queuing theory models of human-machine interaction characterize the operator as a single-channel processor sharing attentional resources serially among a variety of tasks. The human is conceptualized as a "server" processing multiple tasks and "server utilization" or "busyness" is used as a measure of workload. These models generally apply to situations in which performance times are critical. Within queuing theory, performance times include both the time it takes to execute various tasks, as well as the time that tasks must wait before being performed. Rouse (1980) provides a good discussion of queuing theory and its application to human-machine modeling.

The emphasis in queuing models is more on when tasks are performed rather than how they are performed. As indicated by Rouse, these models are most appropriate in multitask situations in which the operator must cope with task priorities and with performance requirements that vary among the tasks. Using Jahns' (1973) categorization of workload (Chapter 2), queuing theory models are concerned primarily with the input load to the operator. A benefit of queuing models is that fractional attention is computed as a function of time and system performance dynamics are taken into account.

The queuing theory approach to workload estimation is generally considered in conjunction with Senders' analysis of monitoring tasks (e.g., Senders, Elkind, Grignetti, & Smallwood, 1966; Senders & Posner, 1976). However, others such as Schmidt (1978), analyzing the workload of air traffic controllers, and Walden and Rouse (1978), modeling pilot decision behavior, have also successfully applied this approach.

Other Mathematical Models

The above sections have suggested the major applications of mathematical models to predicting workload. However, a variety of other modeling approaches have been proposed, but have had limited use in a workload context. For example, Moray (1976) discussed the use of Signal Detection Theory. Signal detection involves asking a subject to detect signals imbedded in noise. Detection of a signal when present is a true positive (correct) and detection of a signal when none was presented is a false positive, (error). By varying the probability of a signal actually present, it is possible to generate receiver operating curves (ROC) which indicate both true signal detection and subject bias for false positives. Signal Detection analogues have been developed and used within optimal control theory (Levison & Tanner, 1971); this application may be useful for predicting OWL.

Finally, White, MacKinnon and Lyman (1985) have outlined a model based on a modified Petri net system for workload estimation and prediction. The work was an attempt to demonstrate that the model was sensitive to workload manipulations and achieved promising results. However, the predictive capability of the model is still to be demonstrated.

Summary

The application of manual control theory to workload estimation and prediction is generally restricted to environments involving continuous controlling tasks. During that period when workload was practically synonymous with vehicular control, manual control models were easily the most interesting and promising class of techniques providing predictions to system designers. In the present day, these models may be adapted to estimate measures generally associated with OWL, but the mathematical sophistication required to develop or even understand the models limits their applicability. Detailed system parameters must also be provided to exercise these models fully; these parameters are frequently not available during early concept development. Consequently, manual control models are generally not viable for most conceptual system evaluations.

The popularity of mathematical models seems to have waned. Information theory was most popular in the 1960's and manual control theory and queuing theory predominated during the 1970's. Although many of these models have experienced considerable success within the domain for which they were intended, they seem to have been supplanted in the 1980's by computerized task analysis and simulation models. A major problem with mathematical modeling is the absence of explicitly defined workload parameters. Thus, while model outputs may identify and quantify particularly busy periods within a given time slice, or particularly high periods of information transfer, it is never quite clear how, or if, these phenomena relate to high workload. This observation, it should be pointed out, is not restricted to mathematical models alone and probably has relevance to most analytical techniques and methodologies.

There is always a place for a useful mathematical model, even if the model is not as broad as one would like. An obvious and hopeful evolution would be that certain of these mathematical models, especially the optimal control model which can cover aspects of queuing formulations, might be incorporated into the simulation models. It would certainly seem feasible to bring such models into simulations in a form which more people could use.

Task Analysis

Task analysis techniques have a long history (Drury et al., 1987) and are the most commonly used of all analytical tools for predicting workload in the preliminary design process. This is partly due to the military requirement for a task analysis to be performed during system development (MIL-H-46855B). It is a fairly natural extension from this requirement to derive OWL estimates from the task analysis.

Task analysis methods seek to produce operator performance requirements as a function of fixed increments of time defined against a scenario background. The basic task analysis process begins with definition of a mission scenario or profile. Next, the general mission requirements are systematically decomposed into mission segments, functions, and operator tasks; the tasks in turn are decomposed into detailed operator task element requirements. These elemental task requirements are defined as operator actions required to complete the task within the context of the system characteristics. Thus, the timing and sequencing of operator actions will depend on the nature and layout of controls and displays. The result of the analysis is an operator activity profile as a function of mission time and segment, essentially a time-based analysis of performance requirements.

Workload Definition

A natural consequence of time-based task analysis is to define OWL operationally as time stress. Time stress is expressed as a ratio of Time required (Tr) to perform a task over the Time available (Ta), yielding Tr/Ta. Workload situations of concern are, therefore, those which cause the operator to approach the edges of the performance envelope, that is Tr/Ta approaches 1.0. This definition encompasses only one aspect of workload: time stress. A technique incorporating such a definition is useful, but probably best utilized as an initial coarse filter to identify gross design deficiencies and for cases in which the time required for a task is well defined. Diagnosticity, in the time-line technique, is limited to identifying general functional limitations where demands exceed operator capacity to respond within some time frame.

Other approaches are more detailed in the analysis of tasks, further partitioning them into components relevant to sensory channel or body part (e.g., eyes, ear, hand, foot, etc.). Recent methods have included a still more detailed analysis structure in an attempt to identify types of cognitive loads imposed on the operator. However, these more detailed approaches still typically contain time stress (Tr/Ta) as a major contributor in the estimation of workload. Nevertheless, diagnosticity improves by virtue of identification of specific components that may be overloaded.

There are many variations on the basic task analysis structure. The differences will be clarified in the discussions of each of the methods. The models presented here are intended to be illustrative of the class of information that can be integrated into the models and the nature of the results that can be obtained. A review of many task analysis techniques may be found in Meister (1985).

Time-Based Task Analysis Procedures

Timeline Task Analysis. A recent application of the timeline analysis technique employing the Tr/Ta metric is that described in Stone, Gulick and Gabriel (1987). They used this technique to identify workload with respect to specific sensory-motor channels encountered in overall aircraft operations. Validation efforts are reported by the authors, with the results indicating that the procedure "...provides a reasonably accurate index for predicting the time required to complete observable tasks within the constraints of an actual mission."

Workload Assessment Model (WAM). The Workload Assessment Model was introduced as part of a more comprehensive human-machine system design aid, Computer Aided Function-Allocation Evaluation System (CAFES). WAM is intended to estimate the effects of alternate function allocations on OWL (Edwards, Curnow, & Ostrand, 1977). In WAM, a mission timeline is developed which indicates what tasks are performed during the mission and in what sequence they are performed. The individual sensory-motor channels (e.g., eyes, ears, hands, feet, etc.) that are involved in the execution of each task are identified. WAM computes the channel utilization percentage including the amount of time that each channel is occupied within a specific time segment. Percentages over a specified threshold level are considered excessive, and identify either function allocation deficiencies, design inadequacies, or both.

A variant of WAM, the Statistical Workload Assessment Model (SWAM), allows shifting excessive workload tasks in time in an attempt to reduce the workload level. This, in effect, is a rescheduling of tasks to reduce time stress. Linton, Jahns, and Chatelier (1977) report one application of SWAM. They examined a conceptual VF/VA-V/STOL aircraft to determine whether a single pilot could manage the aircraft and its avionics subsystems in defined mission phases. The results indicated the potential single-pilot operability for the aircraft, but did not establish any validity measures for the assessment technique.

The Time-Based Analysis of Significant Coordinated Operations (TASCO). TASCO analyzes tactical mission cockpit workload using the standard time-based approach (Roberts & Crites, 1985; Ellison & Roberts, 1985). The basic analytical component of the method is the EDAM (Evaluation, Decision, Action, and Monitoring) loop. Evaluation takes into account the impact of information display design. The decision is made by the pilot based on training, experience, tactical doctrine and situational awareness applied to the evaluation of the data displayed. The decision results in an action via the cockpit controls which is then monitored to evaluate the outcome of the action.

Two types of analysis are performed in TASCO. The first is crewstation task analysis, which is a design evaluation performed by an SME using a 5 point rating scale to judge design elements that are especially crucial to mission performance. The second is a Busy Rate Index analysis, which is essentially a Tr/Ta estimate over a set time interval. How the above mentioned EDAM loops are integrated into these analyses is unclear, as is the current state of development of the TASCO model.

Computerized Rapid Analysis of Workload (CRAWL). CRAWL involves expert opinion superimposed upon a task analysis background with two basic sets of inputs (Bateman & Thompson, 1986; Thompson & Bateman, 1986). The first set of inputs includes task descriptions generated by SMEs on the proposed system under study, along with SME-generated workload ratings for four separate channels - visual, auditory, cognitive and psychomotor. Additionally, the average time for task completion and a short verbal description of each task are included. The second set of inputs contains timing information, including the starting time for each occurrence of each task executed during the mission segment. Overall workload for each time segment is computed by summing the workload ratings for the four channels.

In an effort to validate CRAWL, workload estimates obtained while operators flew a single seat simulator were compared to CRAWL predictions of workload for six combat mission scenarios. Overall, an average correlation of 0.74 was found between the predicted workload levels and pilot subjective workload ratings obtained during the simulation study. The correlation indicates good agreement between the two measures.

Workload Index (W/INDEX). W/INDEX combines mission, task, and timeline analyses with theories of attention and human performance to predict attentional demands in a crewstation (North, 1986). It differs from other task analytical techniques by providing estimates of the effect of time-sharing loads imposed by concurrent task demands. W/INDEX estimates workload demands for one-second segments based on individual task difficulty and time-sharing deficits.

W/INDEX operates on the following data:

- Crewstation interface channels,
- Human activity list,
- Attention involvement levels,
- Interface conflict matrix, and
- Operator activity timelines.

W/INDEX was applied to three different conceptual cockpit designs and was demonstrated to be sensitive to design changes although apparently not validated against empirical studies.

The McCracken-Aldrich Approach

McCracken, Aldrich, and their associates have recently developed a task analysis approach for predicting OWL that does not rely solely on the time-based definition of workload (McCracken & Aldrich, 1984; Aldrich, Craddock & McCracken, 1984; Aldrich & Szabo, 1986). These authors attempted to improve the diagnosticity of workload predictions by identifying four (and later, five) behavioral dimensions which contribute to overall workload levels. They were also among the first to isolate explicitly cognitive workload demands. This approach has impacted other task analysis methods (e.g., CRAWL described above) and simulation methods (e.g., Micro Saint, described below).

The McCracken-Aldrich methodology involves performing mission and task analyses that generate a rough timeline (i.e., one without a strict time scale) of operator tasks. These tasks are further partitioned into elemental task requirements which, based on system characteristics, are used to generate estimates of workload for up to five workload dimensions (Szabo et al., 1987):

- cognitive,
- visual,
- auditory,
- kinesthetic, and

- psychomotor.

Workload assessments are made by assigning numerical ratings for each of the applicable workload components. These ratings represent the difficulty or effort associated with performing the task. It is in the ratings that this technique differs most from other task analyses. The ratings are generated by comparing verbal descriptors of the task elements with the verbal anchors identified with each scale value. The five workload components are assigned scale values of one through seven (Szabo et al., 1987). For example, during the post mission checklist of a helicopter, the copilot performs the task of inspecting the exterior of the aircraft. That task, in turn, requires that the copilot "visually inspect each side of the airframe" (visual scale value = 2) and "evaluate the current status of the airframe for damage" (cognitive scale value = 2). The scale and verbal anchors for the cognitive component are presented for illustrative purposes in Table 3-3.

Estimates of the duration of each task element ultimately are developed to construct a detailed task timeline using one-half second intervals. Total workload is estimated by summing across concurrent task elements for each workload component, visual, auditory, cognitive, kinesthetic, and psychomotor, during each time interval. If this sum exceeds a threshold value, e.g., 7 on visual, then the operator is assumed to be overloaded on the component. The frequency of overloaded intervals for each mission segment can then be determined and the causative workload component identified.

Table 3-3. Cognitive workload component scale (McCracken & Aldrich, 1984).

Scale Value	Verbal Anchors
1	Automatic, simple association
2	Sign/signal recognition
3	Alternative selection
4	Encoding/decoding, recall
5	Formulation of plans
6	Evaluation, judgement
7	Estimation, calculation, conversion

Hamilton and Harper (1984) proposed a modification of the McCracken-Aldrich technique. Their variant replaces the summation method of workload estimation with an interference matrix approach for detailed workload analysis. This matrix defines acceptable, marginal, and unacceptable workload levels for each of four workload components. A series of decision rules are then employed to define whether or not entire mission segments have acceptable, marginal, or unacceptable workload levels. This technique alleviates certain interpretive problems concerning the implication of having, for example, a total mission segment rating of 10 on visual tasks with a scale range of only one to seven. Validation efforts with this technique indicated that it is sensitive to task differences and reflected empirical pilot opinion ratings obtained in simulation studies. It was also found to predict slightly higher workload ratings than those obtained by the empirical rating; this bias may be desirable for design purposes.

Cognitive Task Analysis

The idea that a more detailed task-analysis structure can provide increased diagnosticity is an important one. Combining this idea with the fact of increased influence of cognitive tasking leads to the approach of detailed decomposition of cognitive workload into component types. This approach has been developed and applied to selected aircraft systems (Zachary, 1981). As in more traditional task analysis, operator tasks are decomposed and are grouped into four primary categories: cognitive, psychomotor, motor, and communicative/interactional. A mission scenario is independently developed with a variable timeline grain depending on mission segment (for example, an attack mission segment may be decomposed to second by second events whereas a return-to-base segment may be decomposed into five minute intervals). Operational personnel then work with cognitive scientists to map operator tasks onto the scenario timeline. Next, workload levels are assigned to each operator task as the scenario unfolds. Workload ratings for the same task may vary depending on the mission segment in which it is performed.

In particular, the workload analysis is based on a set of workload rating scales that describe five distinct types of cognitive workload:

- planning difficulty,
- prediction difficulty,
- calculation difficulty,
- information processing complexity, and
- information absorption complexity.

In addition, eight other workload scales are utilized in the categories of: psychomotor (pointer movement and writing), motor (button-pushing frequency and keyset entry frequency), and interactional (interruption frequency, interruption magnitude, communication frequency, and communication complexity).

Applications of this methodology for each time segment yields individual ratings on thirteen scales and averaged ratings for the four categories (cognitive, motor, psychomotor, and interactional), as well as an overall workload (average of 13 measures). This promising methodology has been recently applied to two systems – the P-3C anti-submarine warfare tactical coordinator (Zaklad, Deimler, Iavecchia, & Stokes, 1982) and the F/A-18 single-seat aircraft (Zachary, Zaklad, & Davis, 1987). Little formal validation has as yet been accomplished, although the effort is still ongoing.

Summary

Task analysis has demonstrated high utility. The definitions of workload within the various task analyses are not complete, but being based principally on time, they are clearly closely related to perceived OWL. Indeed, the criteria for most tactical missions contain a temporal component in the measure of effectiveness (MOE). And it is true, if a task cannot be done within the time requirements, of what importance is accuracy? For those situations in which time required (Tr) is estimated to be near or approaching the performance envelope boundaries (Ta), additional evaluations can and should be performed to identify OWL components which may be adversely affecting performance time.

Simulation Models

The application of simulation models to the workload estimation problem is conceptually an extension of the traditional operator-in-the-loop simulation procedure. The major difference, of course, is that the simulation effort is expanded to include a simulated operator. Similarly, simulation may be considered an extension of task analysis. Within simulation models, differences among the models include: (a) whether operator characteristics must be defined along with system and environmental characteristics or (b) whether the operator model is included as part of the overall simulation model. Meister (1985) and Chubb, Laughery and Pritsker (1987) review simulation models and their applications.

Good descriptions of the operator, system and operational environment are the first prerequisites. Given such a model, the problem remains to define an appropriate workload index that can be used to compare differences across tactical missions, system configurations or operational uses. In most instances, a task loading index such as time required/time available is used. Furthermore, some simulation models can predict not only operator workload, which itself may or may not affect system performance, but also system performance for future comparison with empirical measures of effectiveness (MOEs).

Simulation vs. Task Analysis

The distinction between the task analysis methods and the computer simulation methods is not always clear. Simulation models have been described as elaborated task analysis methods with consideration of the statistical nature of constituent elements. Most computer simulation models employ a task analysis as part of the development effort, and most task analytical methods are now computerized. The basic distinction that is intended in this categorization is that the task analysis methods produce operator performance requirements as a function of fixed increments of time defined against a scenario background. Simulation models, in contrast, attempt to represent (simulate) operator behavior statistically within the system under study and produce measures of effectiveness for human-system performance. In other words, running a computerized task analysis twice would yield identical answers. Running a simulation model twice would not necessarily yield the same results due to different consequences of branching statements and statistical modification of task times and, where appropriate, performance accuracies.

Types of Models

Recently, Sticha (1987) has discussed two general types of models to simulate human performance. According to Sticha, the difference between these two existing classes can be stated in terms of the ways in which the control of sequencing of the behaviors is accomplished. The first of these is a network model. This approach controls the order directly in a network by means of the way the analyst has developed the procedures – order is defined in the procedure. Network models are a combination and amalgamation of a number of techniques: flowcharts, program evaluation and review technique (PERT), Markov models, decision trees, and reliability models. The second method of simulation is the production rule approach. Production models control the ordering through a set of production rules and through these rules by the environment. Sequencing is indirectly inferred by a set of rules which associate a behavioral action with an environmental event. The actions are performed only when the environmental conditions of the rules have been met – order is thus defined by the environment. There are no true production models used in workload, however, there are several hybrid models employing both the network and the production rule approach. Sequiturs Workload Analysis System (SWAS) and the Human Operator Simulator (HOS) are examples of hybrid models. Although both classes of models may in some situations produce identical results, they have different capabilities. In particular, Sticha points out that procedural tasks are characterized by internal control whereas tasks involving the recall and application of rules are driven by the environment.

Siegel-Wolf Network Models

The majority of simulation models are derivatives of the network model developed by Siegel and Wolf (1969). Siegel and Wolf models come in several variants involving the number of operators simulated. The basic purpose of the models is to provide an indication to developers about where in a proposed system the operators may be over-stressed or under-stressed. The models predict task completion times and probabilities of successful task completion. The variable that relates to workload is termed stress. Stress is caused by:

- failing behind in time on task sequence performance,
- a realization that the operator's partner is not performing adequately,
- the inability to successfully complete a task on the first attempt with the possible need for repeated attempts, or
- the need to wait for equipment reactions.

Both time and quantity of tasks enters into the stress definition. Note, however, that task quantity can be reduced to time. Stress is typically calculated as the ratio of the sum of the average task execution times to the total time available. A task difficulty factor has been included in recent model developments (Meister, 1985).

Input to the network model typically consists of 11 data items for each subtask and operator (Meister, 1985). These are shown in Table 3-4. There are many sources of the necessary data, including detailed task analysis, but the major source is direct questioning of subject matter experts (SMEs). The type of data input is usually not sensitive to design changes within a specific type of system component (e.g., dials), but can differentiate between different types of components (e.g., dials vs. status lights). Model outputs include a number of performance measures such as number of runs, average run time, number and percent of successful runs, average, peak, and final stress, and several others. The primary uses for these models are for the coarse prediction of system effectiveness and design analysis. Siegel-Wolf models are typically used for discrete task modeling.

SAINT/Micro SAINT

An important extension of the Siegel-Wolf model is called the System Analysis of Integrated Networks of Tasks (SAINT). SAINT, along with its microcomputer version Micro SAINT, is actually a task network simulation language. It contains a number of process branching rules, multiple distributions for modeling individual task operations, and a Monte Carlo sampling procedure for determining task execution. As a general purpose simulation language, it provides a framework and contains little implicit information toward a developed model. This means that operator, system, and environmental characteristics must be entered by the modeler. Micro SAINT provides a menu-driven interface to facilitate this development effort. SAINT's underlying approach to estimating workload is the same as the Siegel-Wolf models. SAINT defines stress as the ratio of time required to complete a task to the time available (Tr/Ta). SAINT can be used to model both discrete and continuous tasks.

Table 3-4. The eleven data elements required for each subtask and operator for Siegel-Wolf Models (from Meister, 1985, p. 125).

1. Decision subtasks,
2. Non-essential subtasks,
3. Subtasks which must be completed before it can be attempted by another operator,
4. Time before which a subtask cannot be started,
5. The subtask that must be performed next,
6. Average task duration in seconds,
8. Average standard deviation of task duration,
9. Probability of being successful,
10. Time required for all remaining essential tasks, and
11. Time required for all remaining non-essential tasks.

Micro SAINT has been used in conjunction with a separate workload estimation methodology. Laughery et al. (1986) used Micro SAINT to predict OWL in four alternative helicopter cockpit designs using a model which incorporated characteristics of the operator, a helicopter, and the threat environment as task networks. OWL was assessed during the Micro SAINT simulation following the technique developed by McCracken and Aldrich (1984). The use of the McCracken-Aldrich task analysis required the assignment of workload requirements for each of five workload components – auditory, visual, cognitive, kinesthetic, and psychomotor dimensions – for each operator activity. Thus, each task is characterized by its requirements for each of the components. Overall, workload could then be assessed for tasks executed individually or in combination if executed concurrently. Workload was assessed at 2-second intervals in order to track it through the simulated mission scenario. The results demonstrated that the methodology was sensitive to variations among the helicopter designs, and that specific components overloads could be identified. The authors report that total development and execution time was on the order of 10 weeks, although subsequent development times can be substantially less. This integration of network simulation with more robust and diagnostic workload prediction methodologies is a promising development.

Simulation for Workload Assessment and Manning (SIMWAM)

Another related simulation methodology is called the Simulation for Workload Assessment and Manning (SIMWAM) (Kirkpatrick, Malone & Andrews, 1984). SIMWAM is based on SAINT and the Workload Assessment Model (WAM) (Edwards, Curnow, & Ostrand, 1977), but has been developed to make it especially suitable for examining manpower issues, as well as individual operator workload, in complex multi-operator systems. SIMWAM has been used to assess workload and manpower issues for an aircraft carrier's aircraft operations management system (Malone, Kirkpatrick & Kopp, 1986). Specifically, the SIMWAM application focused on the effects of incorporating an automated status board (ASTAB) into the existing system. The scenario involved 35 shipboard operators engaged in the launch-recovery cycle of 25 aircraft. Two workload assessments were made: one on the existing baseline system and another with the proposed ASTAB. The results of the analysis indicated that the introduction of ASTAB would allow a reduction in the number of required personnel by four individuals. That conclusion was based on the workload having been reduced to near zero for these four individuals, where workload was defined by number of tasks they performed and the amount of time that they were busy (i.e., occupied with tasks). Also, the number of operators who were heavily loaded (i.e., busy at least 75% of the time) was reduced by one half. Thus, SIMWAM provides a basis for predicting the impact on manpower requirements of proposed system modifications. Such results are especially meaningful to program managers.

Sequiturs Workload Analysis System (SWAS)

Sequiturs Workload Analysis System is a hybrid model incorporating features of both types of models, network and production techniques, as discussed in the introductory section on simulation models. (Holley & Parks, 1987). In contrast to the network models discussed above which are performance simulation tools, this model has been developed specifically for workload analysis. The definition of workload is the by now familiar time required over time available (Tr/Ta). Success is defined strictly in terms of the Tr/Ta ratio.

SWAS contains a structured helicopter task database, organized according to task categories which in turn are broken into task blocks containing task elements. (This task analysis follows requirements in MIL-H-46855B.) Each task element in the database has ten attributes including the mean time and standard deviation, and differentiation of discrete and continuous tasks. It also has built in assumptions about the organization and functioning of behavior, following the Wickens (1984) resource model. This model plays a major role in the organization, sequencing, and resource time-sharing of task elements as well as modification of performance times. (See Navon [1984] for a critical review of the resource model.)

Additionally, SWAS contains a Methods Time Measurement (MTM) module which is used to assist the user in producing mean performance times. Finally, equations are built in to adjust for types of clothing and individual differences (on a scale from 1 = good to 9 = bad). Both means and standard deviations are adjusted in a multiplicative manner in the equations.

The model has received several validation studies at Bell Helicopter comparing the simulation results with results from operator-in-the-loop studies using both simulation and actual flight of a single pilot helicopter. In these studies, error rates predicted by SWAS differed from operator times by 1% to 8% (underestimate).

Human Operator Simulator (HOS)

The Human Operator Simulator (HOS) is a simulation model using an approach different from the Siegel-Wolf models (Wherry, 1969; Lane, Strieb, Glenn, & Wherry, 1981; Harris, Glenn, Iavecchia, & Zaklad, 1986). The original HOS approach was based on four assumptions:

- Human behavior is predictable and goal oriented, especially for trained operators.
- Human behavior can be defined as a sequence of discrete micro-events, which can be aggregated to explain task performance.
- Humans can time-share (switch) among several concurrently executing tasks.
- Fully trained operators rarely make errors or forget procedures.

The implication of these assumptions is that the model is deterministic, that is, the outcomes of operator actions are derived from functional relationships formed as equations rather than by sampling from a probability distribution.

The latest version, HOS-IV, is a general purpose simulation facility that provides the capability to predict system performance by dynamic, interactive simulation of the human operator, the hardware/ software system, and the environment. HOS-IV is implemented on a microcomputer (IBM PC-AT) (Harris, Iavecchia, Ross, & Shaffer, 1987). HOS-IV contains an enhanced user interface to assist in defining, executing, and analyzing the simulation. The HOS-IV user can build independent models of the environment, hardware, and operator to the desired level of detail using a top-down approach. Operator task times can be crudely estimated and entered into the simulation or tasks can be decomposed in order to utilize the set of basic human performance micromodels resident in HOS. For example, a target recognition task could be modeled coarsely by merely specifying a time estimate for the overall recognition process. Alternatively, the recognition task could be decomposed into micro-events such as an eye movement followed by a visual perception followed by a decision. In the latter case, HOS-IV would determine the time required to complete the task.

HOS-IV contains a library of human performance micromodels that can be used to simulate the timing and accuracy of particular human behaviors. The core set of micromodels are all based on experimental literature and can be accessed by the user. The micromodels include: eye movement, visual perception, decision time, short-term memory, listening and speaking, fine-grained control manipulation, hand movement, and walking. These micromodels can be easily modified or replaced entirely.

Models of environment, system, and operator are defined with the following simulation building blocks:

- An object database containing names and characteristics of the entities to be simulated (for example, Emitters, Radar, Displays, and Controls).
- A set of rules which start an action when conditions are appropriate.
- A set of sequential actions required to accomplish a process. The process can be defined for the environment, system, or operator. Operator processes can utilize human micromodels provided by HOS-IV.
- An optional set of events which define external occurrences that affect the simulation flow at predetermined times.

The result of the simulation is a detailed timeline of operator, hardware, and environmental events and actions which can be summarized and analyzed for a broad variety of purposes. Standard output analyses are available which provide statistics associated with performing tasks, subtasks, and basic behaviors. This includes the number of times a micromodel is executed, the mean and standard deviation of the time to complete a process, and the percent of simulation time spent on each process. Additionally, the user can define and access information on system measures of effectiveness.

Lane et al. (1981) identified a number of applications and validation efforts over a wide range of systems. Generally, the results have been very favorable. HOS allows a very detailed model to be developed, providing a greater degree of diagnosticity than other simulation models. HOS is probably more applicable as a follow-on analysis after less detailed analytical techniques have been used to refine the system design.

Model Human Processor (MHP)

Card, Moran and Newell (1983, 1986) have developed a potentially powerful collection of micromodels collectively called the Model Human Processor (MHP). Via the MHP, they have established a framework for presenting data contained in the human performance literature in a manner which will make it more accessible to those involved in the engineering design process. They partition human behavior models according to their application to the perceptual, cognitive, or motor systems, and focus on simpler, more widely applicable models that capture the predominant characteristics of a problem. Models such as these can be used to define limits of operator-system effectiveness to any scope required. The MHP micromodels are currently only described in the literature. Some of the MHP models, however, have been directly incorporated into the HOS library and are accessible to simulation modelers. Further work in the development and application of human performance models is required. MHP has proven a fruitful model for analysis of computer interfaces, not covered by other models (Card, Moran & Newell, 1983).

Summary

In recent years, a number of new simulation tools have been developed. Simulations offer a unique opportunity to evaluate both time and accuracy of performance. There is a cost, however, for gaining the accuracy evaluation and that is the additional time required for developing the simulation. However, this may be a small price to pay in the context of overall system development costs.

For the most part, more user friendly versions of simulation models have been developed in the last several years. As additional modules and computer tools are developed and more complete databases are built, simulation techniques will move to the forefront of analytical workload techniques.

Overall Summary and Concluding Comments

Analytic techniques can be used to make predictive workload assessments early in system development. An important characteristic of these techniques is that they may be used before there is an "operator-in-the-loop." Therefore, workload predictions may be available to have impact on early system design.

Analytic techniques can be divided into five major categories:

- Comparison,
- Expert Opinion,
- Mathematical Models,
- Task Analysis Methods, and
- Simulation Models.

This analytic technique taxonomy provides a useful structure in which to classify workload assessment tools that can be used while system concepts and alternatives are being explored. The categories of techniques described require different information and specific techniques may be more appropriate for answering different kinds of questions.

Some of the analytic techniques have not yet been systematically formalized or fully validated (e.g., comparison). Further work should be done to develop these techniques for workload assessment that can be used very early in conceptual development and system design.

CHAPTER 4. EMPIRICAL TECHNIQUES - PRIMARY TASK MEASURES

Overview of Empirical Techniques

With this chapter, we begin the review of empirical techniques used to measure operator workload. As discussed in Chapter 2 and illustrated in Table 2-1, we divided empirical techniques into four major categories:

- Primary Tasks,
- Subjective Methods,
- Secondary Tasks, and
- Physiological Techniques.

Each of these major classes of measures has been researched extensively and we have provided an overview of a number of studies in each category. Further, each of the categories has distinctive features, especially in the context of workload definitions, and these features are manifested in the literature. We have sought to capture these distinctions and differing flavors in our reviews, and accordingly the review for each category differs both in the approach to the literature and organization.

In our discussion of OWL assessment techniques, the overall intention is both an analysis, especially in the context of sensitivity and diagnosticity, and an integration of the literature. The objective of this integration is to provide practical guidance for designers, developers, and evaluators of systems. It is recognized that the individual who should be concerned with human workload issues cannot wade through hundreds of studies to obtain OWL assessment guidance. Resources are very limited and should be expended largely performing the OWL assessment, not learning about workload research. Thus, each class of OWL techniques is reviewed with summaries and recommendations provided.

A Summary Evaluation of Empirical Techniques

Because of the amount and variety of material to follow, a summary evaluation of selected techniques is shown in Table 4-1. The entries in the table represent the authors' considered judgments on the sensitivity, cost and effort, and diagnosticity for a number of the techniques to be discussed. The techniques shown in Table 4-1 were judged on a basis relative to all the other measurement techniques, not just within their own category. Also, the techniques were rated independently for each of the three criteria. (The authors of this volume have collectively worked with virtually every technique in the table.) Please note, it may be better to use a technique rated 'Low' than no technique at all. Although relative judgments have been attached to these techniques, all techniques can all be used to obtain information regarding OWL. In addition, as a point made throughout this report, multiple measures of workload should be used to obtain more complete information regarding potential and existing OWL problems.

Table 4-1. Summary of empirical techniques judged for sensitivity, cost, and diagnosticity.

Technique	Sensitivity	Cost/Effort Requirements	Diagnosticity
Primary Task Measurements			
System Response	Low/High[1]	Low Cost Moderate Effort	Low
Operator Response	High[1]	Low/Moderate Moderate Effort	Moderate/High
Subjective Methods			
Analytic Hierarchy Process	High[2]	Low Cost Low Effort	Moderate[2]
Bedford	High[2]	Low Cost Low Effort	Low
Cooper-Harper	High for psychomotor	Low Cost Low Effort	Low
Modified Cooper-Harper	High	Low Cost Low Effort	Low
NASA-TLX	High	Low Cost Low Effort	Moderate/High
SWAT	High	Low Cost Low Effort	Moderate/High
Psychometric Techniques	High	Low Cost Low Effort	Low
Interviews	Varies	Low Cost Low Effort	Moderate/High
Questionnaires	Varies	Low Cost Low Effort	Moderate/High

[1] Varies with workload

[2] Represents some uncertainty about sensitivity and diagnosticity due to limited research.

Table 4-1. Summary of empirical techniques judged for sensitivity, cost, and diagnosticity (Cont.).

Technique	Sensitivity	Cost/Effort Requirements	Diagnosticity
Secondary Tasks			
Embedded Secondary Task	High	Low Cost Low Effort	Moderate/High
Choice Reaction Time	Moderate	Moderate Cost Low Effort	Moderate
Sternberg Memory Task	Moderate	Moderate Cost Low Effort	Moderate
Time Estimation Task	Moderate	Moderate Cost Low Effort	Moderate
Physiological Techniques			
Blink Rate	Low	Moderate Cost Moderate Effort	Low
Body Fluid Analysis	Low	Low Cost Low Effort	Low
Evoked Potentials	Moderate	High Cost High Effort	High[3]
Eye Movements & Scanning	High	High Cost High Effort	High[3]
Heart Rate	Moderate	Moderate Cost Moderate Effort	Moderate
Heart Rate Variability	Moderate	Moderate Cost Moderate Effort	Moderate
Pupil Measures	Moderate	High Cost Moderate Effort	Moderate[3]

[3] The rating applies within a narrow, specialized range.

The sensitivity rating reflects the relative ability of the measure to discriminate among different levels of workload. The cost and effort requirements reflect a judgment of the overall resource requirements including personnel, time, effort, and equipment. The diagnosticity reflects the usefulness of the measure in pinpointing the processes involved in high workload.

As will be seen below, primary task measurement has some interesting properties that cause sensitivity to vary with workload. Although relative judgments have been made regarding secondary tasks, there is uncertainty as to their sensitivity and diagnosticity outside of the aviation environments. For those entries with more than one rating, the judgments are intended to reflect the range of sensitivity or diagnosticity. There are several areas where insufficient information exists to make a judgment, although preliminary findings suggest the degree of sensitivity or diagnosticity; these uncertainties are marked. A few entries reflect the variable nature of the measurement technique depending on specific situations; these are also marked.

Video recording of operator performance is a useful tool in OWL assessments, but can not be easily placed in a table such as the one presented. It can be used as an important, practical empirical method and should not be overlooked in developing empirical measurement procedures.

For primary task techniques, there are a very large number of specific measures that have been used -- nearly every situation requires its own measures. Because of this diversity, theoretical and conceptual analysis is very important. First, we have classified primary task measuros into system performance and operator performance measures. Then, the development and selection of unique primary measurements is considered. Primary task measurement is covered in this chapter.

Subjective methods research is very different. The emphasis is on assessing the operator's experiences and the amount of subjective effort expended. Most OWL research is concerned with subjective rating scales, but there are a relatively small number of these scales in wide use. Accordingly, our review focuses on these scales in detail and analyzes the comparative features of the rating scales. Subjective techniques are covered in Chapter 5.

For secondary task techniques, the situation is somewhat similar to primary task techniques in that a great many individual measures have been used; however, a substantial part of the research has utilized a limited number of techniques. In our discussion, we have examined some underlying theoretical issues to the use of secondary tasks. These issues revolve around attempts to assess residual capacity or to fill and load that residual capacity. Our review and analysis takes the point of view that real systems are multitask environments and that the secondary task paradigm is most effective in that context. Chapter 6 covers the secondary techniques.

Finally, physiological techniques represent a different class of techniques. In particular, these techniques generally require specialized expertise, extensive equipment, and procedures sometimes difficult to perform outside the laboratory. We have provided some background and rationale for the use of these techniques, techniques that probably assess activation or arousal. Chapter 7 covers the physiological class of techniques.

Primary Task Measurement

The goal of system development is to produce a system which reliably achieves its mission. The operator is an important part of the system. System performance is a combination of operator performance and the hardware system and is reflected in meeting the mission goals. It is the operator's task by means of decision making, integration of information, manipulation of controls, etc. to guide the hardware toward successful completion of the mission. The system responds to the operator's commands. Thus, it is reasonable to talk about two kinds of performance: the operator and the system. A statement about operator performance is meaningless unless system performance is also acceptable. Accordingly, there is a need to measure both.

OWL, as was discussed in Chapter 2, is not the same as performance of the operator or the system. OWL was defined as the relative capacity to respond. OWL arises as the interaction between the operator and other system components during mission execution. Workload evaluation assesses this interaction, i.e., the contribution of the operator to the system and the impact of the hardware and other situational components on the operator. Stated differently, workload evaluation assesses the location of the operator within the workload envelope. One approach to assessing OWL is by means of primary task measures.

Primary Task Definition

Even though it may seem surprising, it is not always clear what is meant by a primary task. In flying or driving, the primary task for the operator is to keep the rubber side down, that is, operate the vehicle in a manner that will maintain proper vehicle orientation. But the operator may have other important functions within a mission. Communications is often considered a secondary or subsidiary task. However, if an aircraft is performing a scout function, accurate and timely communication would be of utmost importance to the mission. Similarly, what is a copilot's primary task? In some helicopters, the job is designated as Copilot/Gunner. At least by the designation, the Copilot/Gunner has two primary tasks, and he may also have to handle communications. Thus, here and in other easily developed examples, the operator may have several primary tasks.

During the course of a mission the emphasis on various operator tasks will change, that is, the priorities associated with one primary task will change in relation to another. What is labeled as the primary task may change depending on the specific situation and where the operator is in the overall mission. For example, most investigators analyzing a mission have seen the need to divide the mission into segments to capture the flavor of the different task emphases and priorities. Further, the label, primary task, is sometimes simply a definition assigned by interested analysts, such as a workload evaluator. (This definitional issue will arise again in Chapter 6 when consideration of secondary tasks is discussed.) In some cases, the definition is clear, in other cases however, it may seem somewhat arbitrary. For these reasons, it is better to think of multiple tasks rather than a single primary task. In short, one needs to evaluate several tasks coupled with the priorities associated with those tasks and not just a single primary task. We are going to discuss primary tasks in a general way and as though the definition were always clear.

Primary task measures are important as part of a battery of workload measures. However, it should be pointed out that this position is not universal. Some authors (e.g., Hart, 1986a; O'Donnell & Eggemeier, 1986) state that primary task measurement may not be useful in workload assessment. Certainly, there are many examples where this is true. However, the very fact that the results appear to be contradictory suggests that further analysis and clarification is in order. The following discussion provides the necessary clarification.

Primary Task Measurement Types

Primary task techniques may be categorized into two broad types. Type 1 includes those measures which are of the system and contain a contribution in some form (sometimes unknown) of operator performance. For an instrument landing task, glide slope and localizer errors, often measured using root mean square (RMS) or standard deviation of relative position, are of this type (Wierwille, et al., 1985). Type 2 measures, by contrast, are a more direct index of operator performance, often finer-grain, fine structure measures that reflect strategies adopted by the operator to cope with task demands. For the landing task example, this could be the number of control movements per unit time (as measured at the stick or column, not at the aircraft control surfaces).

To understand the importance and implications of this classification, let us consider again the relationship between performance and workload shown earlier in Figure 2-1. Figure 4-1 is a replot of Figure 2-1.

- Region 1 – the operator's load is too low. (This region is not discussed in this report).
- Region 2 – the operator's load is not excessive, because additional resources can and may be mustered to maintain performance, and the performance level is held relatively constant and high.
- Region 3 – the operator's load has become excessive. In this region, the load increases well beyond the operator's capability for compensation, and performance levels become asymptotically low.

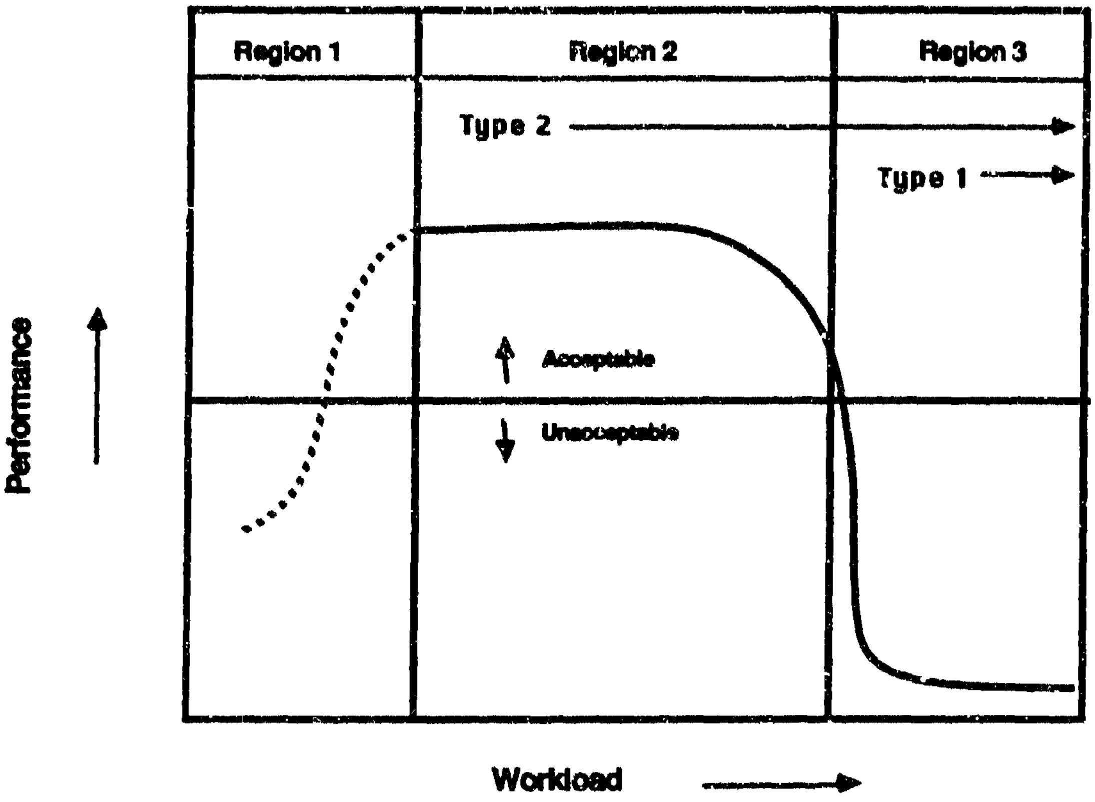


Figure 4.1. Hypothetical effect of workload on sensitivity of Type 1 and Type 2 measures.

These response regions provide a framework for understanding where Type 1 and Type 2 primary task measures are sensitive to workload. Sensitivity is a critical issue with primary task measurements. Type 1 will be sensitive in Region 3, whereas, Type 2 will be more sensitive in Region 2 as well as covering Region 3.

Type 1 Measures: The System+Operator. Type 1 measures of primary task performance are indices of system+operator performance. Typically, they include measures of human tracking errors or other measures of system performance (e.g., Wierwille et al., 1985). However, measures of system performance such as engine thrust, RPM, movement of control surfaces could be classified as a Type 1 measurement, since changes in thrust, for example, reflect operator acdvities plus system lags. Similarly, any measure of effectiveness (MOE) for mission performance would ordinarily qualify as a Type 1 measure. Type 1 measures were an initial focus of workload research, no doubt because of their association with the quality of system performance (Sanders, 1979; Williges & Wierwille, 1979). This category of measures provides an index of system performance (MOEs) and is useful in this regard.

Type 2 Measures: The Operator. Type 2 measures of primary task performance are defined here as those which assess the nature of operator performance directly (Hart, 1986a). The measurement can take several different forms: a measure may be directed at quantity, frequency, or quality criteria of operator performance. Type 2 measures may also be directed toward detecting the fine structure of operator performance, i.e., those that link operator activity to measurable performance (Hart, 1986a). In general, the category includes such measures as: (a) control movements per second in a psychomotor task, (b) response times in a perceptual or cognitive task, (c) errors of omission, (d) errors of commission, or (e) communications response times in a communications task (Wierwille et al., 1985). The very reason Type 1 measures are insensitive is also the reason Type 2 measures are sensitive: As the operator copes with workload and under increasing load marshalls greater resources to hold Type 1 performance constant, the operator may perform differently and patterns of performance may change and fine structure tends to shift. Type 2 measures are valuable because these shifts may provide evidence of a change in OWL and hence provide a means to assess OWL levels.

Comparison of Type 1 and Type 2. Several studies have used Type 1 measurements in parallel with Type 2 measures. For example, O'Donnell and Eggemeier (1986) discuss a study by Schultz, Newell and Whitbeck (1970) which showed increases in turbulence had no effect on glide slope error (Type 1). Similarly, Wierwille et al. (1985) did not find significant effects of task loading on glide slope error. However, if one examines the frequency of control inputs for wheel, column, or throttle (Type 2), there is a clear effect of turbulence on frequency of control movements in the Wierwille et al. study and in two other separate studies by Dick (Dick, 1980; Dick et al., 1976). (In the Dick studies, pilot ratings of handling quality ranged from 3 in a no turbulence condition to 7 in a high turbulence condition and there was no effect of turbulence on glide slope error.)

Sanders, Burden, Simmons, Lees, and Kimball (1978) tested nine helicopter pilots on each of three levels of stabilization augmentation (for yaw, pitch and roll) during hover. Altitude control was under manual control for all three conditions. Thus, the stabilization device should facilitate altitude control since less effort would be expended on the other dimensions. Type 1 measures did not show any effect for the three levels of stabilization augmentation or for altitude control. In short, system performance did not differ for the three conditions. By contrast, Type 2 measures for Fore-Aft control and pedal movements for altitude control showed significant variations with both fewer movements and smaller magnitude of movement with the stabilization augmentation devices operational. Other Type 2 measures did not show an effect. (Of interest, average pilot ratings only ranged from 3.1 to 4.3 for the three conditions, showing a relatively small subjective spread among the conditions.) In accordance with the Type 1 - Type 2 distinction, Type 1 measures were insensitive while Type 2 measures were sensitive to variations in variables that affect workload.

Summary. Although we have cited only a few studies, the general statement can be made that Type 1 measures of system+operator are not often sensitive to workload manipulations, however, they are important in system evaluation considerations. Type 2 measures of the operator directly generally show effects on relevant dimensions; relevant dimensions being those measures one would reasonably expect to show a difference as in the Sanders et al. (1978) study. Type 2 measures are essential for workload evaluation.

Enhancing Type 2 Measures: The Fine Structure of Behavior

Some investigators have questioned both the sensitivity and the diagnostic capability of primary measures. However, as shown above, when one makes the distinction between Type 1 and Type 2 measures, it is clear that Type 2 primary measures are sensitive. Furthermore, it is possible to enhance sensitivity and often diagnosticity by examining the fine structure of behavior. This enhancement can be an especially valuable approach, because time and money are almost always limited. Some ideas are developed below which provide background for practical application of such measures.

Many of the primary task measures shown to be especially sensitive to workload variations are indicators of strategy shifts (Hart,1986a). While some investigators have avoided the strategy interpretation, the results reported seem to be consistent with the ideas being developed here. (See O'Donnell and Eggemeier [1986] for additional studies in this category.) *Strategy* is widely used in describing behavior and the term without restriction encompasses too many types of action descriptions including style, S-R mapping process, etc. Accordingly, we will use *rule* as a more neutral, easily defined, and precise term. This usage here has parallels with the idea of rules in production models of behavior (Card, Moran, & Newell,1983). Rule driven performance changes also proved to be sensitive to manipulations of load for many of the techniques and measures investigated by Wierwille et al. (1985).

A brief digression. In order to draw out the value of Type 2 measures fully, it is appropriate to consider what is meant by *rule driven* performance. In some sense, one could argue that all behavior is rule driven. There are global rules which might involve survival, for example, and there are more detailed rules. What we are interested in primarily is the fine structure of rule driven behavior. Identification of rules is done by inference from detailed examination of the performance measures. Accordingly, measurement should be done with care so as to permit the correct inferences to be made.

A hypothetical simple visual discrimination task will illustrate a detailed example of rule driven behavior. In this experiment, the stimuli are purposely picked and will be either an H or an N, since they differ only in the cross bar. They will be presented at the same point in space (and all other conditions are controlled). When the H is presented, the operator is to push the left button to indicate his response; when the N is presented the operator will press the right button. The experiment is performed, but the operator is not informed about the fact that his reaction time is also being recorded in addition to accuracy. After completing the data collection for this task, the operator is asked to do the experiment again. For this second case, however, the operator is told about recording reaction time AND the operator is told to respond as fast as possible. To continue our hypothetical example, after having finished the second task, the operator is asked to do it yet a third time. The operator is told that the performance on the second case contained too many errors and consequently the operator should be more accurate, but reaction time will again be measured in this third case.

What would the collective results of such an experiment show? First, the average response time would be different for each of the three experiments. The second case would be the fastest, the third case the next fastest, and the first case the slowest. Second, the accuracy would also differ with Case 1 best and Case 2 worst. Specifically, the pattern of performance is different for the three cases. Why? Basically, because the operator was working under three different sets of instructions or three sets of rules. These rules might be, in order of execution the cases above:

Case 1. Do the visual discrimination as accurately as possible: Time is not a factor.

Case 2. Do the discrimination as fast as possible: Accuracy is less important than time.

Case 3. Do the discrimination as fast and accurately as possible: Accuracy and time are of equal importance.

The time-accuracy tradeoff is a well known phenomenon in the reaction time literature (e.g., Posner, 1978; 1986). However, there is a catch - not all people use all of the rules or in the manner logic would dictate. Unless the instructions are explicit, any of the three rules may be used depending on the individual. Only when the conditions are changed by instruction or by a situational demand, is it possible to determine which of the rules were used, that is, several levels of a workload variable need to be included. A further requisite to discovering the rules is the measurement of two components of behavior, time and accuracy. Without both measures, the discovery of this underlying rule structure would be difficult. Additionally, had we measured it, we might have found that the force applied to the response buttons differed for the three cases as well. This, and other measures of behavior, could provide additional information about the fine structure of operator behavior.

Application of Fine Structure Measures. The process of identifying fine-structure and rule-related measures for Army systems may be illustrated with a hypothetical example (patterned after the communication task experiments of Wierwille et al. [1985]). This communication task approach has been shown to be sensitive to workload manipulations in another context (Green & Flux, 1977). These rules will be executed according to perceived time demands. That is, an assumption is made that the operator will not perform at a rate higher than the situation demands. These hypothetical rules might be:

Rule 1.	If	TIME IS AVAILABLE	then	DO NORMAL communications pace.
Rule 2.	If	TIME IS SHORT	then	SPEED UP speech rate

Rule 3. If TIME IS CRITICAL then SHORTEN messages.

Rule 4. If TIME IS VERY CRITICAL then DELAY or ELIMINATE non-essential communications.

Application of these rules by the operator would have important implications on performance and simultaneously reflect changes in workload. The types of changes in performance one would expect for each of the rules are as follows:

Rule 1 performance. Normally paced performance. This forms a baseline against which to compare other performance characteristics under other rules.

Rule 2 performance. Quicker response than Rule 1 and/or message compacted into a shorter time. A few errors of omission.

Rule 3 performance. Fewer words in message than Rule 2, rate of speaking similar to performance with Rule 2. Possible errors of omission.

Rule 4 performance. Few words in message and spoken fast as compared with Rule 3, less essential messages omitted or delayed. Both errors of commission and omission.

Some Previous Studies. Many tasks can be dissected in this way and anticipated operator performance rules established. The exact types of performance rules and associated changes will differ with the situation. For instance, pilots will make more control movements on the stick (and/or wheel) and possibly throttle under heavy turbulence conditions than under light turbulence conditions. However, just because one observes a change in the performance measure, one cannot necessarily conclude that workload is higher. Use of an autopilot with manual throttle, is certainly a lower workload condition as compared with total manual control; nevertheless, the number of throttle changes increases substantially under the autopilot mode (Dick, 1980). Indeed, this difference represents a rule driven performance change, but not one caused by increased workload; the overall pattern of performance measures is needed to identify the reasons for this finding. Similarly, a reduction in performance may reflect fatigue more than workload per se (Angus & Heslgrave, 1983).

Bainbridge (1974; 1978) has reviewed and discussed performance rules and their role in determining performance. For example, air traffic controllers were asked to find conflicts between aircraft. The controllers used two methods: some controllers arranged flights under their control geographically and others by altitude. Those controllers who used the altitude approach were able to perform better (faster and more efficiently) than those who used geography (Leplat & Bissert [1965] cited in Bainbridge, 1974). Similarly, as the number of aircraft increased, there was increasing simplicity and decreasing redundancy in messages (Sperandio [1971] cited in Bainbridge, 1974). These and other examples are indicative of performance rule changes as a function of task demands and task loading.

Summary. Some Type 2 measures are conducive to identifying rules through the fine structure of performance and others are not, that is, the measures vary in sensitivity for rule detection and identification. There is no easily categorized structure of behavior which fits this fine structure analysis approach. Generally, as in identifying the speed-accuracy trade-off, it is necessary to employ several different measures. Accordingly, one should attempt to tread the fine line between missing an important parameter and burying the analyst in a flood of less relevant data (Hart 1986a). Multiple measures provide greater capabilities for probing aspects of rule driven performance as well as providing potentially enhanced statistical sensitivity via multivariate analysis (O'Donnell & Eggemeier, 1986). Multiple measure [including fine structure] assessment frequently will also serve to overcome the criticism of primary measures as insensitive and non-diagnostic (e.g., Gopher & Donchin, 1986; O'Donnell & Eggemeier, 1986).

Development of Primary Task Measures

A major difficulty with primary measures is their potential lack of transferability across applications (Hicks & Wierwille, 1979; Williges & Wierwille, 1979). The specific measures to be used must typically be developed for each application and may not be used routinely in another application. The difficulty stems from the simple fact that operators may perform different tasks in different systems, and consequently outputs or work products differ. Of course, measures may be and should be adopted across systems in cases where tasks are essentially unchanged (e.g., stick-movements for aircraft evaluations, steering wheel and accelerator movements for driving, communications in a variety of contexts). Because of the potentially reduced transferability of primary measures, general guidance for their development is outlined in the following discussion. Specific consideration is provided for selection, implementation, and preliminary evaluation of reliability of primary measures.

Selecting Measurements on Primary Tasks

Appropriate primary task measures may be devised for each application. Remembering that measures of performance have differing utilities, an investigator should identify measures that are most appropriate for the application at hand. Where appropriate, Type 1 system performance measures may be identified by examining system objectives and outputs. These measures might include number of targets detected, number of targets fired on, accuracy and rate of firing, etc. The measures selected, of course, will be highly dependent on the system under evaluation. For the Type 2 category, usually potential rate and error measures for each task can be identified that provide the requisite direct mapping between operator behavior and measurable performance (Hart, 1986a). In general, latency and error scores are excellent candidates and have been reported as sensitive across a half dozen studies by O'Donnell and Eggemeier (1986).

For identifying Type 2 rule related measures, asking an operator to describe the rules is helpful, but not always useful. There are often differences between what operators do and what they think or say they do (Spady, 1978a). One approach is for the investigator to do a preliminary evaluation before performing the actual test. This can be done by monitoring operator task performances under known varied loads. As load increases, the investigator may then discover the performance rules used by the operator for adaptation or coping with the additional load. Candidate rule-related measures may then be chosen which reflect this adaptation process.

An alternative is to use measures which have been shown to be successful. Meister (1985, pp. 256-263) has considered issues of selecting task measures as well as provided a listing of possible measures for system evaluations based on earlier work (Smode, Gruber, & Ely, 1962). Table 4-2 presents an extract of this listing that may serve as as guide to selecting a variety of primary task measures. In the table, a Category like TIME is a general class of measurement with three Subcategories, e.g., reaction time, which in turn can be applied to several events listed under Description. In the communication example these could include: (a) mean elapsed time between the end of received message and the beginning of the next transmitted message, (b) mean length of each transmitted message, (c) variance in time between end of received message and beginning of corresponding transmitted message, or (d) proportion of messages shed (omitted). The process of identifying rule related measures ultimately involves the mapping of expected rule usage to corresponding measures that reflect the use of such rules.

Errors are an especially interesting measure. Errors can take several forms: omission, commission, or wrong order of execution. Not only might they reflect high workload, they can be the cause of increased workload (Hart, 1986a). When an error is made, often some corrective response has to be made by the operator. This adds on to the number of things the operator has to do and increases time pressure. If the number of errors is substantial, then elimination of the cause of the errors will substantially reduce workload. Any technique that provides diagnosticity, will be of general help.

Table 4-2. Varieties of primary task measure candidates (Meister, 1985).

CATEGORIES	SUBCATEGORIES	DESCRIPTION
TIME	Reaction time, i.e., time to	• perceive event; • initiate movement; • initiate correction; • initiate activity following completion of prior activity; • detect trend of multiple related events.
	Time to complete an activity already in progress	• identify stimulus (discrimination time); • complete message, decision, control adjustment; • reach criterion value.
	Overall (duration) time	• time spent in activity; • percent time on target.
FREQUENCY OF OCCURRENCE	Number of responses per unit, activity, or interval	• control and manipulation responses; • communications; • personnel interactions; • diagnostic checks.
	Number of performance consequences per activity, unit, or interval	• number of errors; • number of out-of-tolerance conditions.
	Number of observing or data gathering responses	• observations; • verbal or written reports; • requests for information.

Table 4-2. Varieties of primary task measure candidates (Meister, 1985) (Cont.).

CATEGORIES	SUBCATEGORIES	DESCRIPTION
ACCURACY		
	Correctness of observation; i.e, accuracy, in	• identifying stimuli internal to system; • identifying stimuli external to system; • estimating distance, direction, speed, time; • detection of stimulus change over time; • detection of trend based on multiple related events; • recognition: signal in noise; • recognition: out-of-tolerance condition.
	Response-output correctness; i.e., accuracy, in	• control positioning or tool usage; • reading displays; • symbol usage, decision making and computing; • response selection among alternatives; • serial response; • tracking; • communicating.
	Error characteristics	• amplitude measures; • frequency measures; • content analysis;. • change over time

Additional guidance for deciding what to measure can be developed through established task taxonomies like the Universal Operator Behaviors that developed by Berliner, Angell, and Shearer (1964). In this organization, human activities in systems are separated into four broad categories:

- Perceptual tasks are sensing tasks; for example, seeing a warning light on an instrument panel.

- Mediational or cognitive tasks are those that involve thinking (e.g., solving mathematical problems).

- Communication includes face-to-face speaking, radio, and other communication tasks.

- Psychomotor processes are manipulative tasks; those which involve muscles or movement (e.g., activating a pushbutton).

By spreading the selection of tasks across these categories, one increases the opportunities for identifying performance measures that are sensitive to workload and are diagnostic of the causes of workload. There is little point, for example, to measure two tasks that fall in the same category. But it would be highly useful to measure two tasks in different categories.

Implementation

Because the operator ordinarily performs the task as part of his duties, primary measures have the advantage that they generally need not be intrusive on the operator nor require specialized training (O'Donnell & Eggemeier, 1986; Williges & Wierwille, 1979). All that is required to obtain Type 1 and Type 2 measures is to instrument the system. There are many instrumentation methods that may be used and their use is dependent on the application. In fielded systems, it may necessary o add sensors as well as transponding or recording equipment. Of course, there are some situations where, due to the absence of ample space for adding instrumentation, there may be a physical space intrusion. This space intrusion is typically less severe than the intrusion required for implementation for some other methods for assessment of OWL (e.g., physiological). Moreover, such limitations can often be overcome with a bit of ingenuity. In simulators, space is usually less of a problem; sensors are often already in place and may be used for the purpose of gathering data on the primary measures. The proliferation and use of microcomputers and interface cards has simplified implementation and reduced space requirements. In general, primary task measures typically take up less space and have fewer implementation difficulties than other methods for assessment of OWL. More importantly, implementation of primary task measures will ordinarily be required for combat system evaluations in the context of MANPRINT considerations.

Reliability of Primary Task Measures

Primary task measures have the potential to provide important information about OWL. This potential will not be realized, however, if the reliabilities of measures across sessions are inadequate. Frequently assessed by correlation coefficients, reliability is the consistency of measurement and involves the accuracy and stability of measures and the observational condition under which the measures are made

(Meister, 1985). Addressing the growing concern with operational performance assessment (General Accounting Office, 1982), Lane and colleagues have recently indicted low reliabilities (and resulting inadequate sensitivities) as a major, chronic problem of such investigations (e.g. Lane, 1986; Lane, Kennedy, & Jones,1986). Their indictment of the operational literature parallels that directed at human performance evaluations as part of environmental investigations (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986). These parallel indictments are based on mathematical arguments as to the limitations on sensitivity imposed by low reliabilities.

Much of the operational literature addressed by Lane and colleagues is concerned with flight performance evaluation; however, they also point to other examples including operator performance in armor (e.g., Biers & Sauer, 1982). The body of evidence points out the need to evaluate reliabilities and sensitivities of primary measures before use in operational performance investigations. Reliabilities and sensitivities may be evaluated using a diverse number of direct and indirect approaches. Three of these approaches are delineated below because of their particular utilities in the context of combat system evaluations.

Operational Test Experience. Measures may be selected based upon sensitivities and reliabilities demonstrated in previous operational tests of similar systems under test conditions generally parallel to those planned. Since many systems are derivatives of previous systems, this approach has the advantage of building upon experience. A possible disadvantage is that of discouraging the development of potentially superior measures. This disadvantage can be overcome by using good, demonstrated measures along with new ones which are specific to issues of interest to the system. For totally new systems, the practitioner may be forced to develop new measures but can build on experience to the extent the functions and missions are similar.

Baseline Assessment or Pilot Test. Reliabilities of measures may be determined by administering parallel operational test conditions to a group of subjects on two or more occasions as part of a baseline evaluation before an operational test formally begins. A pilot test may be an opportunity to obtain the baseline measurements. Although not widely appreciated, the results of such baseline evaluations may be used to (a) evaluate the readiness and training of the operator-subjects, and (b) identify fundamental or gross evaluation problems before resources are wastefully expended. Baseline assessments of reliabilities and other measurement qualities have previously been applied by a number of researchers (e.g., Bittner et al., 1986; Jobe & Banderet, 1986). Averaged correlations across occasions may be used to identify measures with highest reliabilities and initial potential for sensitivity when the numbers of operator-subjects are modest (Dunlap, Silver & Bittner, 1986).

Theoretical Considerations. Reliabilities and sensitivities of measures may occasionally be evaluated based upon theoretical considerations. Where aspects of a single performance may involve a trade-off by an operator (e.g., accuracy and speed in a data input task), a theory-based measure (e.g., transmitted-bits/second) integrating these trade-off aspects may be simpler to consider, more reliable, and sensitive. Care is required before use of such theoretical integrations, however. For example, the signal detection theory sensitivity metric (d') may not be applicable because low frequency of errors (Parasuraman, 1986). In addition to this caveat, there are several related scoring procedures (occasionally advocated to control for individual differences) whose use should be questioned, if not avoided. These include slope, difference, and proportion of baseline procedures which have been attacked for low reliabilities and on other grounds based upon both analytical and empirical results (cf., Bittner et al., 1986, pp.700-701).

Summary. These three reliability considerations are directed at a range of combat system evaluation contexts. The operational test experience approach is applicable where an evaluation history exists and the baseline assessment approach may be applied where some preliminary data can be collected. The scoring consideration approach is applicable when there is neither an appropriate evaluation history nor an opportunity to gather preliminary data. These three reliability approaches appear to span most evaluation contexts.

Overall Summary

Primary task measurements are divided into two categories. Type 1 measures are of the system (including the operator) and are used to establish and verify the meeting of mission goals. Type 2 measures are of the operator directly and are used in the evaluation of OWL. Additionally, when several Type 2 measures are used in combination, it is sometimes possible to assess the fine structure of behavior and determine performance rules; we advocate taking several Type 2 measures in every evaluation. In general, Type 2 primary measures are shown to be sensitive to workload variations while Type 1 measures are not typically sensitive.

Systems differ in their primary task(s) and measures used in one situation may not always be applicable in another situation. Accordingly, general guidelines are laid out for selection of measures and their implementation. A list of possible measures, based on time, accuracy, and frequency is presented. Special consideration is given to reliability of measures and the means to assess the reliability.

CHAPTER 5. SUBJECTIVE METHODS

"If the person *feels* loaded and effortful, he *is* loaded and effortful whatever the behavioral and performance measures show" (Johanssen, Moray, Pew, Rasmussen, Sanders, & Wickens, 1979, p. 105).

"...mental workload should be defined as a person's private subjective experiance of his or her own cognitive effort" (Sheridan, 1980, p. 1).

The primary purpose for the use of subjective methods is to gain access to the experiences of the operator. Physical workload can be observed, but mental workload occurs internally and can only be inferred by observers. Subjective methods seek to obtain and quantify the opinions, judgments, and estimations of the operators. Indeed, some investigators suggest that subjective methods are the most appropriate methods by which to measure workload. For example, when mental workload is defined as "a person's private subjective experience of his or her own cognitive effort" then workload measurement is "best and most directly done by a subjective scale" (Sheridan, 1980, p. 1).

Investigators interested in mental activities have worked on measurement and scaling of judgments. Many mathematical techniques are available to handle subjective opinion; in recent years some of these have been applied to workload. There has been much written on the use of subjective methods for measuring workload. The multitude of reviews indicates quite clearly the attitude of the research community for the extensive use and the value of subjective methods (Gartner & Murphy, 1976; Moray, 1979b, 1982; O'Donnell & Eggemeier, 1986; Wierwille & Williges, 1978, 1980; Williges & Wierwille, 1979). Although some researchers think that subjective reports are of low value, most think these methods can provide significant information about operator workload (Hart, 1986a).

There are many reasons for the widespread use of subjective measures. As outlined by O'Donnell and Eggemeier (1986), these include:

- easy to implement; little (if any) equipment is needed;
- relatively non-intrusive;
- inexpensive (i.e., cost of the measure is low);
- face validity (at the least);
- many good techniques exist; and
- current data suggest they are sensitive to workload variations.

It is important to be familiar with the various subjective techniques currently available and the research that has been performed in their development and refinement Literature describing applications of subjective workload measures give examples of how and for what types of systems subjective workload measures have been used.

The subjective methods can be broken into two broad classes: (a) *rating scales* and (b) *questionnaires and interviews.* Expert opinion might be considered an associated type of subjective measure, but that method of workload assessment was discussed earlier in Chapter 3. Rating scales employ psychometric scaling methods to derive scales with which quantitative estimates of some behavior or characteristic can be made. Questionnaires and interviews rely on written or oral reports and while there may be quantitative aspects to these data, for the most part, the data obtained are qualitative. Rating scales, questionnaires, and interviews have been used extensively in workload assessment. These methods are reviewed and specific subjective techniques for workload assessment are presented, analyzed, and compared in the following sections. At the end of each discussion the technique or method is summarized. Before discussing each technique, an overview of the nature and properties of measurement scales is provided to set the stage for later discussion.

Levels of Measurement and Scales

There is a wide body of information on the use of scales to measure psychological variables. Many ways to create scales have been developed and the resulting scales may have different properties, each may be appropriate for different circumstances. Which method to use depends on the questions that need answering as well as practical considerations. As background to the description and discussion of specific techniques that have been developed and used for operator workload assessment, a brief discussion of various scale characteristics will be presented.

There are four widely-used levels of measurement: nominal, ordinal, interval, and ratio. *Nominal* measures only classify objects and distinguish classes of items. *Ordinal* measures place objects in order of magnitude, although distance between the objects is not defined. For example, on an ordinal measurement scale, a stick with a rank of 4 would not necessarily be twice as long as a stick with a rank of 2. *Interval* levels of measurement possess equal intervals between objects; there is a standard unit of measure but without a fixed zero point, e.g., a thermometer. *Ratio* measures have equal intervals and a known zero point. A ruler is an example of a ratio measurement; a 6-inch stick is twice as long as a 3-inch stick. Ratios can then be formed and statements about the relative amount of a characteristic being measured can be made (Allen & Yen, 1979).

A scale is an organized set of measurements (Allen & Yen, 1979). The different types of scales can be produced by different methods. Scales that list values of a property along a line, even if the properties are placed an equal distance apart on the line, are at least ordinal and may be interval. Just because lines are equidistant on a piece of paper does not mean the scalar is interval. The method of paired comparisons also produces scales with ordinal levels of measurement. Interval scales can be produced through Thurstone's method of comparative judgments or conjoint measurement. Ratio scales can be obtained using methods of estimation where observers effectively make judgments of the ratio between the magnitudes of two perceptions. These methods are described in detail in books on psychological or psychophysical measurement (e.g., Edwards, 1957; Gescheider, 1985). However, it is important to realize that the way in which scales are developed will determine whether a scale has nominal, ordinal, interval or ratio properties. This level of measurement in turn will be one of the major factors in determining how the data can be interpreted.

Another characteristic of a measurement scale is its dimensionality. In essence, this is an indication of what the scale is intended to measure. There can be unidimensional scales that are intended to measure only one aspect or attribute. Multidimensional scales, on the other hand, are intended to measure more than one dimension concurrently. Specific statistical methods are available to create multidimensional scales and these have been used to create scales that specifically address OWL. Whether the scale is uni- or multi- dimensional has implications as to what is to be measured (i.e., what is to be rated) and how workload is conceived. For example, a global, unidimensional rating of workload implies that there is a single attribute of workload that can be identified and rated. For such a rating, operators have to combine internally all aspects of workload into a single metric and the degree to which various aspects contributed to the overall rating are not ascertainable. Tsang and colleagues have employed such a unidimensional overall workload scale using a line divided into 20 intervals with the end points anchored at low and high workload (Tsang & Johnson, 1987; Vidulich & Tsang, 1987). With multidimensional scales it is especially important that the relative importance of the various measured components of workload be identified explicitly. For example, the NASA-Task Load Index (TLX) uses six dimensions while the Subjective Workload Technique (SWAT) uses three.

Operators and observers can both be asked to make ratings. Operators can make judgments about their subjective experiences. At the same time, observers could monitor the behavior of the operators and make judgments about the level of workload the operator is experiencing. This is essentially a subjective opinion about someone else's subjective experience. However, since observers cannot observe internalized operator activities such as information processing and monitoring, their judgments may be less useful than those of the operator. On the other hand, observers may be able to see more than a busy operator (Hart, 1986a). For example, operators may experience tunnel vision, while the observer maintains a larger field of view. The use of ratings of the same tasks or mission segments by both operators and observers may provide more reliable information regarding OWL.

Workload Rating Scales

Subjective scaling techniques have been used to develop rating scales for workload measurement. In general, these rating scales have been developed in aviation communities for measurement of pilot workload, with the exception of the Modified Cooper-Harper scale. In some instances, the rating scale is specific to pilot activities and would need modification to extend its applicability to non-piloting activities. In other instances, the scale would be applicable as it exists to a wide range of tasks and environments. The degree of applicability has been noted where appropriate.

The specific subjective scale techniques described in this report include:

- the Cooper-Harper scale and modified versions of the scale which use a decision tree structure,
- the NASA-Task Load Index and Bipolar scales that obtain individual weighted scores of several dimensions of workload,
- psychometric techniques, such as magnitude estimation and equal-interval scales, and
- the Subjective Workload Assessment Technique (SWAT) which uses conjoint analysis.

Other rating scales which have been developed and used for specific purposes are discussed as examples of applications. Comparisons among the techniques as well as other key issues are discussed following presentation of the techniques.

Cooper-Harper Scale and Variations

Perhaps the most widely used workload-related decision tree rating scale is the Cooper-Harper (CH) scale (Cooper & Harper, 1969). It is a 10-point unidimensional rating scale, resulting in a global rating on an ordinal scale of the experience of piloting an aircraft. It was primarily intended for use by pilots to rate aircraft handling and control qualities, but pilot workload and compensation are mentioned in the scale shown in Figure 5-1. O'Donnell and Eggemeier (1986) discuss the supporting evidence which shows a relation between CH ratings and workload (e.g., Hess, 1977). Wierwille and Connor (1983) also found the CH scale to be sensitive to handling properties. They concluded that the CH scale can be confidently used for tasks that are primarily motor or psychomotor. These findings are generally supported by previous workload literature (Wierwille & Williges, 1980). Haworth, Bivens, and Shively (1986) have recently found a correlation of .75 between CH ratings and NASA Bipolar ratings and a correlation of .79 between CH and SWAT ratings, indicating considerable agreement among the scales and overlap of the underlying psychological dimension.

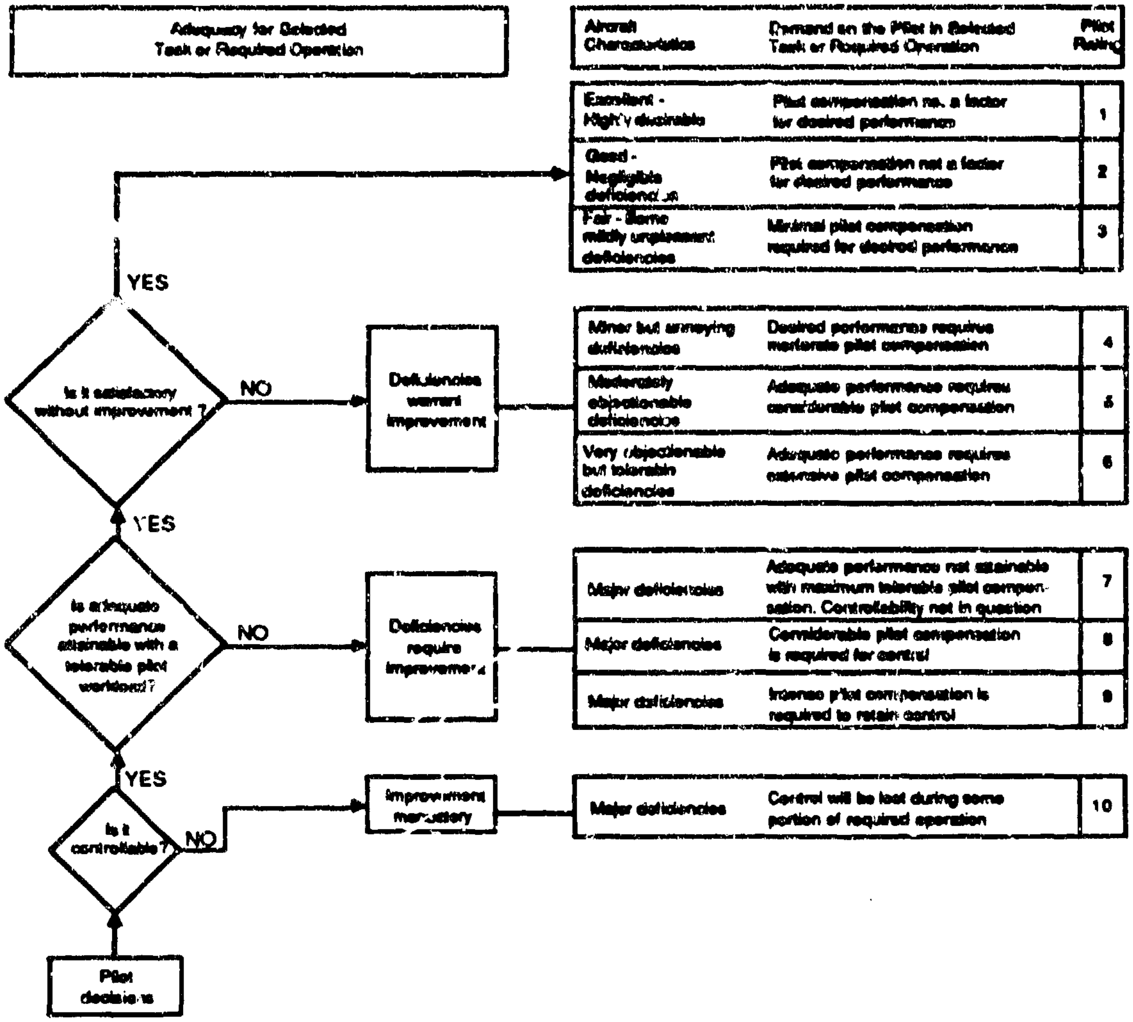
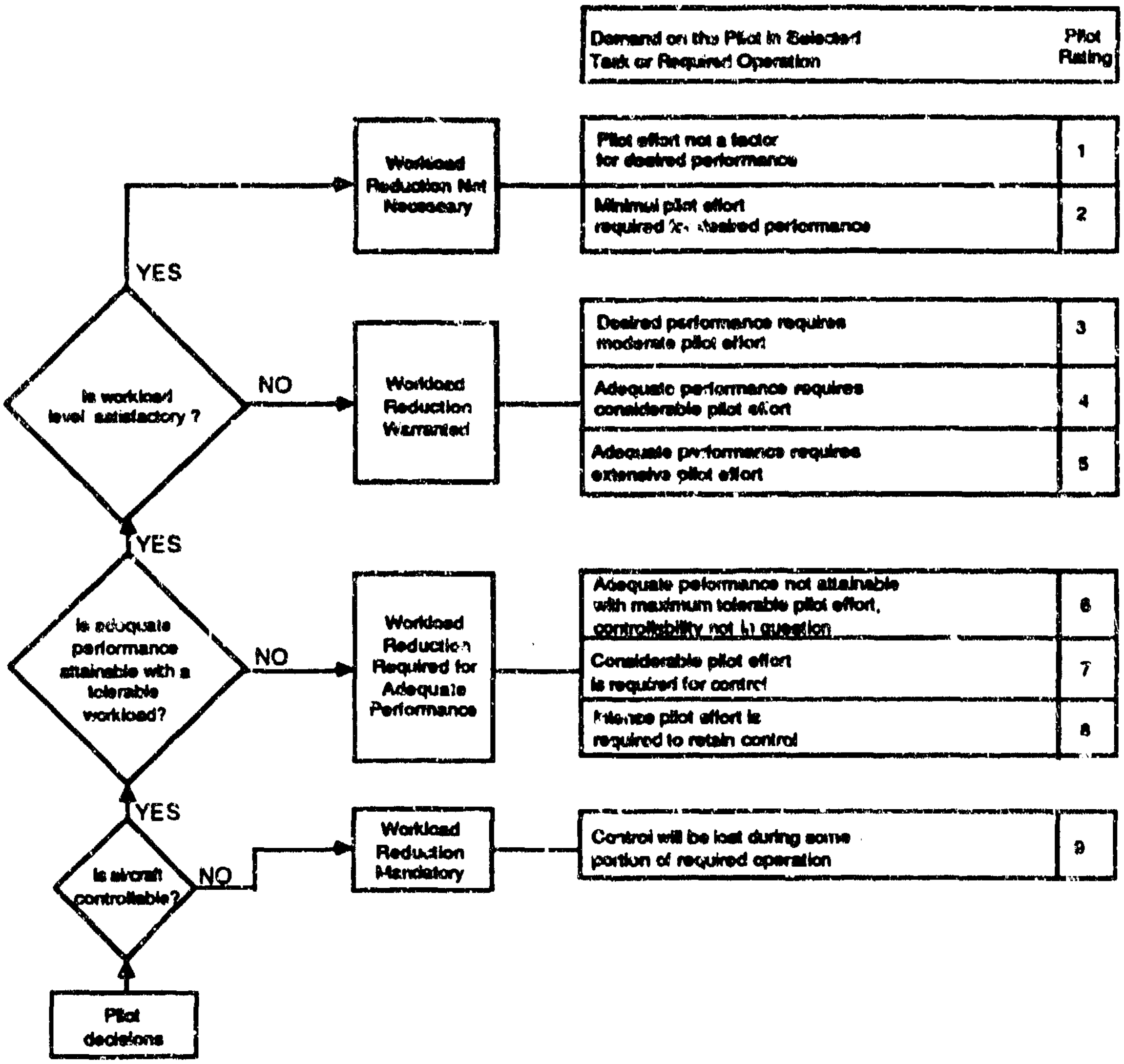

Adequacy for Selected Task or Required Operation
Aircraft Characteristics
Demand on the Pilot in Selected Task or Required Operation
Pilot Rating
Excellent - Highly desirable
Pilot compensation not a factor for desired performance
1
Good - Negligible deficiencies
Pilot compensation not a factor for desired performance
2
Fair - Some mildly unpleasant deficiencies
Minimal pilot compensation required for desired performance
3
YES
Is it satisfactory without improvement ?
NO
Deficiencies warrant improvement
Minor but annoying deficiencies
Desired performance requires moderate pilot compensation
4
Moderately objectionable deficiencies
Adequate performance requires considerable pilot compensation
5
Very objectionable but tolerable deficiencies
Adequate performance requires extensive pilot compensation
6
YES
Is adequate performance attainable with a tolerable pilot workload?
NO
Deficiencies require improvement
Major deficiencies
Adequate performance not attainable with maximum tolerable pilot compensation. Controllability not in question
7
Major deficiencies
Considerable pilot compensation is required for control
8
Major deficiencies
Intense pilot compensation is required to retain control
9
YES
Is it controllable?
NO
Improvement mandatory
Major deficiencies
Control will be lost during some portion of required operation
10
Pilot decisions

Figure 5-1. The Cooper-Harper aircraft handling characteristics scale (Cooper & Harper, 1969).

The Honeywell Cooper-Harper. Other researchers have adapted the decision tree structure used in the Cooper-Harper scale to rating scales for specific use in workload assessment. Wolf (1978) focused on overall task workload rather than aircraft handling qualities in the Honeywell version of the CH scale (Figure 5-2). A comparison of the two scales shows the major differences to be in the use of terms.

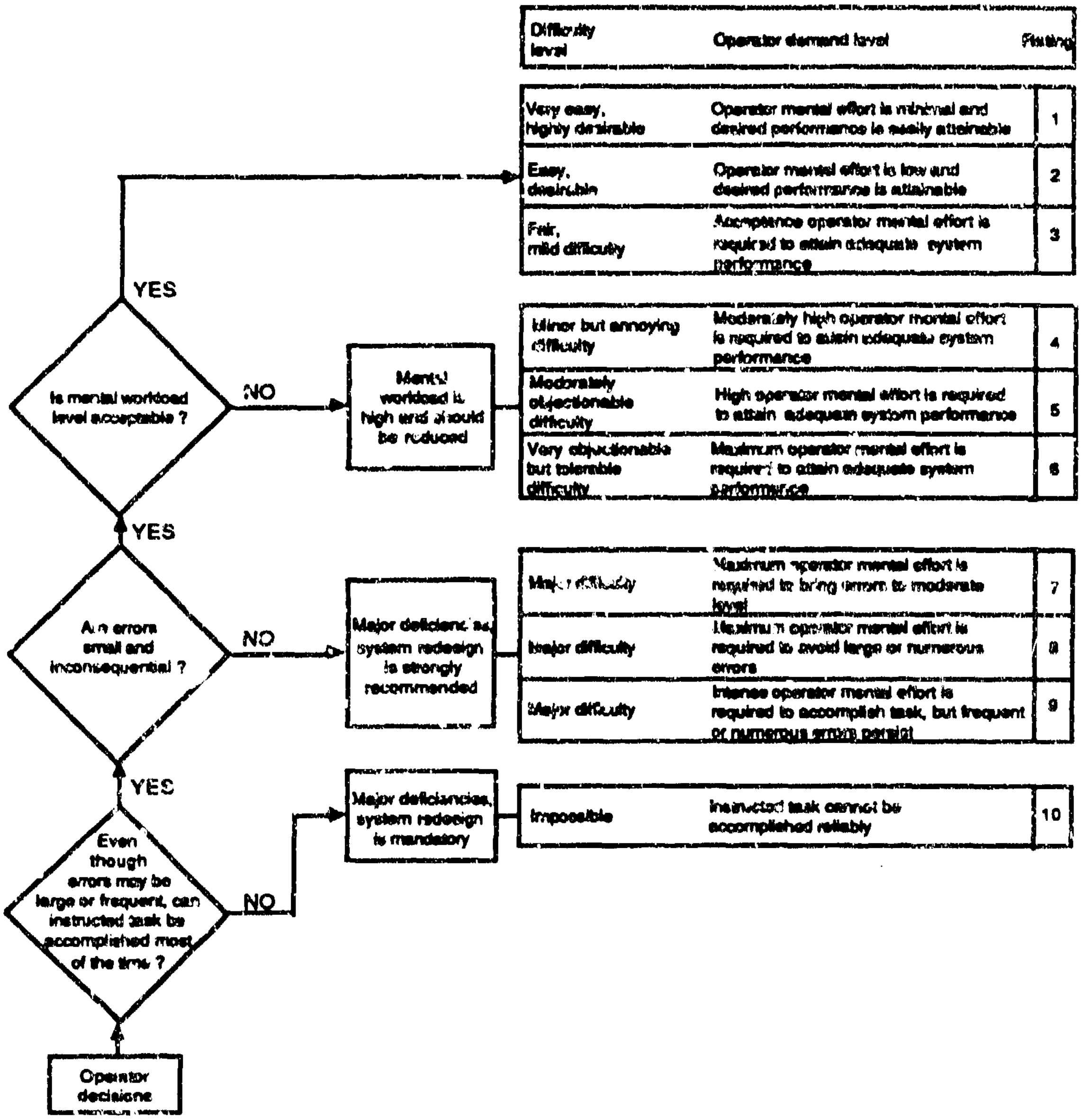
Workload and *effort* in the Honeywell version may be considered task-related, rather than the terms *compensation* and *deficiencies* in the CH, which are more hardware, especially aircraft, oriented. This scale was used in a study of vertical take off and landing (VTOL) aircraft displays (North, Stackhouse, & Graffunder, 1979). In general, the ratings were in agreement with the performance data; however, scores were obtained for only a subset of all conditions. North et al. did not draw strong conclusions concerning the use of this scale because not all factors influencing workload were rated.



Demand on the Pilot in Selected Task or Required Operation
Pilot Rating
Pilot effort not a factor for desired performance
1
Minimal pilot effort required for desired performance
2
Desired performance requires moderate pilot effort
3
Adequate performance requires considerable pilot effort
4
Adequate performance requires extensive pilot effort
5
Adequate peformance not attainable with maximum tolerable pilot effort, controllability not in question
6
Considerable pilot effort is required for control
7
Intense pilot effort is required to retain control
8
Control will be lost during some portion of required operation
9
Workload Reduction Not Necessary
Workload Reduction Warranted
Workload Reduction Required for Adequate Performance
Workload Reduction Mandatory
Is workload level satisfactory ?
Is adequate performance attainable with a tolerable workload?
Is aircraft controllable?
YES
NO
YES
NO
YES
NO
Pilot decisions

Figure 5-2. The Honeywell version of the Cooper-Harper scale.

Modified Cooper-Harper. Wierwille and Casali (1983) developed the Modified Cooper-Harper scale for the purpose of workload assessment in systems where perceptual, mediational and communications activity is present (Figure 5-3). The modification was developed for use in those situations where the task was not primarily motor or psychomotor and the CH might not be appropriate. Wierwille and his colleagues have performed a series of laboratory experiments to validate the Modified CH as a workload assessment technique. Three experiments using this scale are described in Wierwille and Casali (1983). These experiments were performed in a simulated aircraft environment and all were part of larger studies. Six licensed pilots participated as subjects in each experiment. Perceptual tasks



Difficulty level
Operator demand level
Rating
Very easy, highly desirable
Operator mental effort is minimal and desired performance is easily attainable
1
Easy, desirable
Operator mental effort is low and desired performance is attainable
2
Fair, mild difficulty
Acceptable operator mental effort is required to attain adequate system performance
3
Minor but annoying difficulty
Moderately high operator mental effort is required to attain adequate system performance
4
Moderately objectionable difficulty
High operator mental effort is required to attain adequate system performance
5
Very objectionable but tolerable difficulty
Maximum operator mental effort is required to attain adequate system performance
6
Major difficulty
Maximum operator mental effort is required to bring errors to moderate level
7
Major difficulty
Maximum operator mental effort is required to avoid large or numerous errors
8
Major difficulty
Intense operator mental effort is required to accomplish task, but frequent or numerous errors persist
9
Impossible
Instructed task cannot be accomplished reliably
10
YES
Is mental workload level acceptable ?
NO
Mental workload is high and should be reduced
YES
Are errors small and inconsequential ?
NO
Major deficiencies, system redesign is strongly recommended
YES
Even though errors may be large or frequent, can instructed task be accomplished most of the time ?
NO
Major deficiencies, system redesign is mandatory
Operator decisions

Figure 5-3. The Modified Cooper-Harper scale (Wierwille & Casali, 1983).

involved the identification of danger indicators and required a pushbution response. One of three load levels (low, medium, or high) was used for each flight. After each flight, subjects gave a Modified CH rating. The results indicate scores were significantly different for each level of load with the score increasing monotonically with load. The experiment that looked at mediational (cognitive) load used navigation tasks involving various number and complexity of arithmetic and geometric operations for each load level. The navigation solutions were only calculated, not implemented, so the psychomotor elements did not differ. Results showed significant differences between low vs. high and medium vs. high with the score means increasing monotonically with load. The communications experiment involved the use of radio aircraft control and communications tasks including commands for changes in altitude and heading and communications such as reporting call signs and heading, altitude, and airspeed information. Significant differences were found between low vs. medium load and low vs. high load. Score means increased monotonically with load level. The authors conclude that the Modified CH scale ratings are valid and statistically reliable measures of overall workload and that the Modified CH shows a consistent, good level of sensitivity across the three types of tasks (Wierwille, Casali, Connor, & Rahimi, 1985). Modified CH ratings were found to be equally sensitive to task difficulty as SWAT (Warr, Cole, & Reid, 1986).

Wierwille, Skipper and Rieger (1984) conducted two studies to test whether the sensitivity of the Modified CH could be increased by changing from a 10-point to a 15-point scale or by changing the format to computer-based or tabular form. In general, they concluded that the original Modified CH was the most consistently sensitive measure of the five alternatives tested.

The Bedford Scale. The Pilot Workload Rating Scale, also called the Bedford scale, is a decision tree scale derived from the CH. It is shown in Figure 5-4. It was developed by Roscoe and Ellis (Roscoe, 1987a) at the Royal Aircraft Establishment, Bedford, England for workload assessment in the military aviation environment. The technique obtains subjective judgments about workload based on ability to complete tasks and the amount of spare capacity available. It was found that aircrew were able to understand the scale and that it was easy to remember and small enough to be carried on a flight suit knee pad (Lidderdale, 1987).

The Bedford scale has been applied in several workload evaluations of aircrews. Wainwright (1987) reports its application to assess workload for a minimum crew of two pilots of a civilian (BAE 146) aircraft. Three teams of two pilots each participated in the certification program. The evaluation was based on subjective opinion and heart rate monitoring for high workload segments with crews that were asked to fly long duty days with minimum rest. Pilots gave ratings and an observer of the pilot's performance gave a rating as well as the signal to the pilot to rate the previous task. The overall analysis of workload, including subjective measures, heart rate and performance errors, suggested that the two-pilot crews were not overloaded, i.e., crew members reported they had spare capacity.

A similar study measured workload for a two person crew in an advanced combat aircraft during low level maneuvers. Although there was concern that real-time ratings would not be possible, the aircrew were able to give in-flight ratings even in demanding circumstances (Lidderdale, 1987). However, the Bedford scale was found to be inappropriate for obtaining workload assessments during post-flight debriefings.

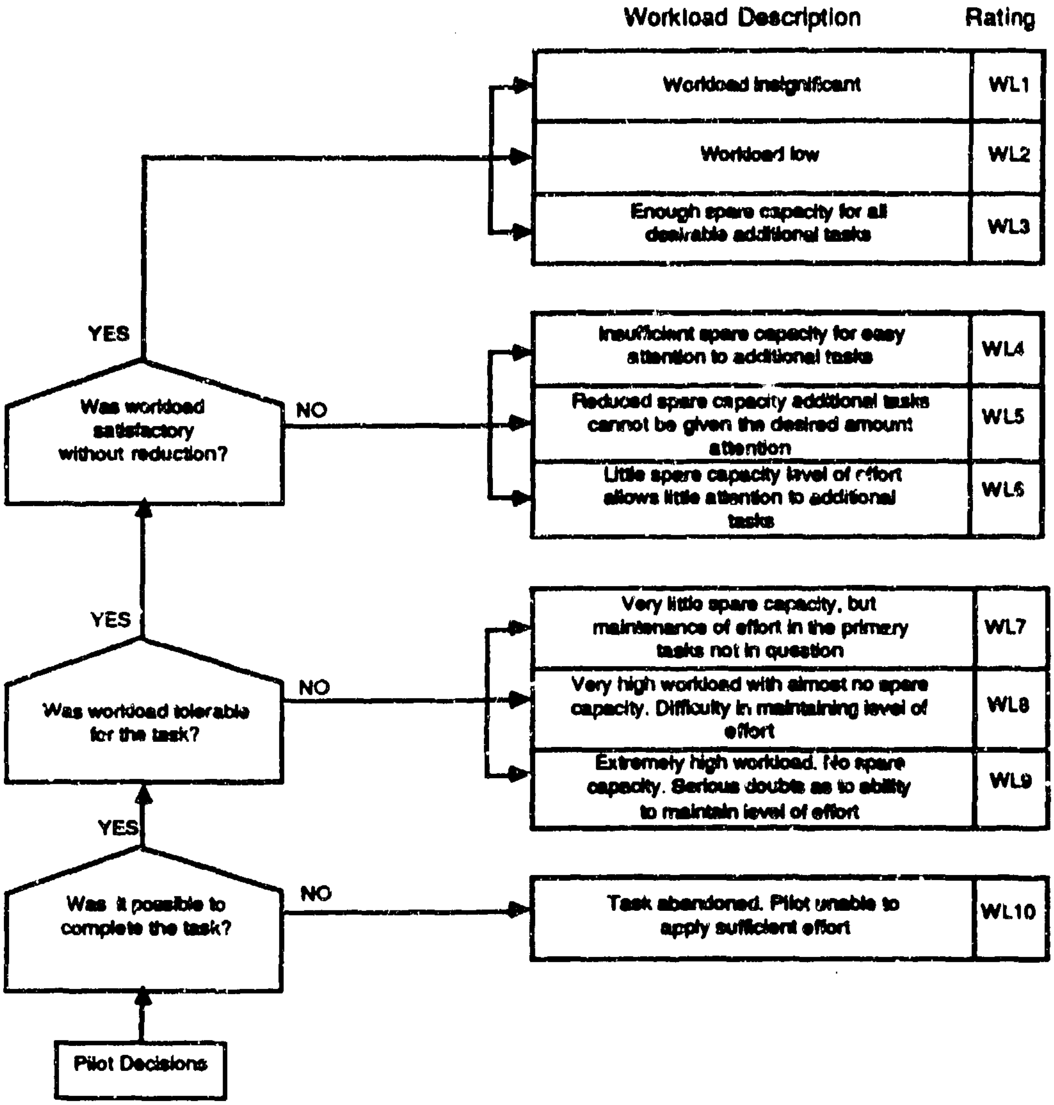

Workload Description
Rating
Workload insignificant
WL1
Workload low
WL2
Enough spare capacity for all desirable additional tasks
WL3
Insufficient spare capacity for easy attention to additional tasks
WL4
Reduced spare capacity additional tasks cannot be given the desired amount attention
WL5
Little spare capacity level of effort allows little attention to additional tasks
WL6
Very little spare capacity, but maintenance of effort in the primary tasks not in question
WL7
Very high workload with almost no spare capacity. Difficulty in maintaining level of effort
WL8
Extremely high workload. No spare capacity. Serious doubts as to ability to maintain level of effort
WL9
Task abandoned. Pilot unable to apply sufficient effort
WL10
YES
Was workload satisfactory without reduction?
NO
YES
Was workload tolerable for the task?
NO
YES
Was it possible to complete the task?
NO
Pilot Decisions

Figure 5-4. The Bedford scale (described in Roscoe, 1987a).

The aircrew found it difficult to reconstruct the complex experiences of the flight and thus they could not be confident in the accuracy of their responses. No other discussion was made of this point, so it is not clear whether the post-flight rating difficulty was due to workload descriptions in the Bedford scale itself or a more general problem that would occur in all post-flight ratings of workload.

The use of the Bedford scale was well accepted by aircrews (Lidderdale, 1987), particularly when tasks are short and well defined (Roscoe, 1987a). Rating pads with 10 push buttons were used as the means to obtain the ratings. Roscoe has identified some limitations in the scale's use: the ratings given are not absolute values and are dependent on the operator's personal experience, therefore comparisons between operators are not valid. Also, real-time ratings may not be be possible if a second person is not available. Like other versions of the CH scale, the Bedford scale produces ordinal data and therefore statistical analysis is limited to rank order tests.

Bedford ratings were found to correlate well with heart rate, although not always consistently (e.g., Wainwright, 1907). The use of the Bedford scale has been primarily in applied settings -- only one study was found where it was used in a controlled setting with defined levels of task difficulty. Tsang and Johnson (1987) used the Bedford scale, the NASA-TLX, and an overall workload scale to measure subjective workload in several manual and semi-automated tasks. The Bedford scale ratings were slightly different from those obtained with the other two measures although the authors suggest these findings support the ability of the Bedford to measure spare capacity. However, these findings are based on a small amount of data and should be used cautiously.

Summary. Workload rating scales based on decision tree structures have been found to be sensitive to different levels of workload in various task types (e.g., Wierwille et al., 1985). The scales have been found to be easy to administer and well accepted by operators. These rating scales have been used almost exclusively in aviation research to assess pilot workload; however, the Modified CH and the Bedford scales would be applicable to other operational environments (with minor modifications such as changing the word *pilot* to *operator*). Finally, interpretations other than as ordinal scales should be approached with great caution because of the nature of the scales.

NASA-Ames Workload Rating Scales

The Human Performance Group at the NASA-Ames research facility has been extensively involved in workload assessment research. As part of the overall effort, much work has gone into the development of workload rating scales as subjective measurement techniques. Two major theoretical considerations influenced the scale development. The first consideration was the multidimensional nature of workload, resulting in multiple workload dimensions. The second consideration was the individual nature of which dimensions of workload are more important for individual operators rating specific tasks. This consideration led to development of individual weighting procedures.

The *NASA-Bipolar scales* are a group of nine scales that reflect nine dimensions of workload plus an overall workload scale. The descriptions of the ten scale dimensions are presented in Table 5-1. Each scale is presented as a single line broken into 20 spaces as shown in Figure 5-5. The operator marks the location on the scale that corresponds to his or her subjective experience related to a specific task. A score from 0 to 100 is obtained for each scale (assigned to the nearest 5). The ratings are assumed to have interval properties. The weighting procedure used to combine individual scale ratings involves a paired comparison task using all pairs of individual dimensions. Paired comparisons require the operator

Table 5-1. NASA Bipolar rating scale descriptions (Hart & Staveland, 1987).

Title	Endpoints	Descriptions
OVERALL WORKLOAD	Low, High	The total workload associated with the task considering all sources and components.
TASK DIFFICULTY	Low, High	Whether the task was easy demanding, simple or complex, exacting or forgiving.
TIME PRESSURE	None, Rushed	The amount of pressure you felt due to the rate at which the task elements occurred. Was the task slow and leisurely or rapid and frantic.
PERFORMANCE	Perfect, Failure	How successful you think you were in doing what we asked you to do and how satisfied you were with what you accomplished.
MENTAL/SENSORY EFFORT	None, Impossible	The amount of mental and/or perceptual activity that was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.).
PHYSICAL EFFORT	None, Impossible	The amount of physical activity that was required (e.g., pushing, pulling, turning, controlling, activating, etc.).
FRUSTRATION LEVEL	Fulfilled, Exasperated	How insecure, discouraged, irritated, and annoyed versus secure, gratified, content, and complacent you felt.
STRESS LEVEL	Relaxed, Tense	How anxious, worried, uptight, and harassed or calm, tranquil, placid, and relaxed you felt.
FATIGUE	Exhausted, Alert	How tired, weary, worn out, and exhausted or fresh, vigorous, and energetic you felt.
ACTIVITY TYPE	Skill Based, Rule Based, Knowledge Based	The degree to which the task required mindless reaction to well-learned routines or required the application of known rules or required problem solving and decision making.

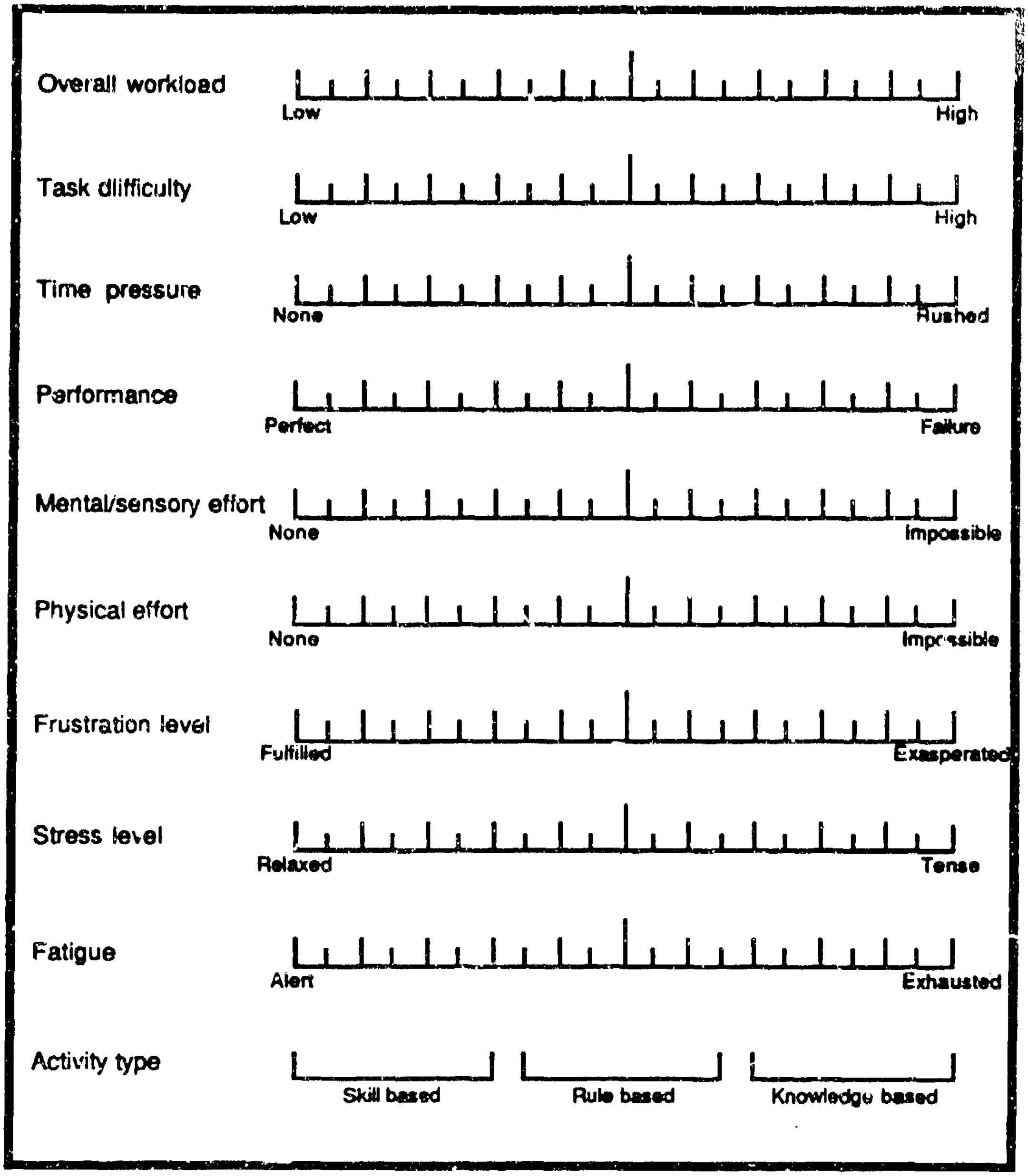


Figure 5-5. The NASA Bipolar rating scales (adapted from Bortolussi, Kantowitz, & Hart, 1986).

to choose which dimension is more relevant to workload for a particular task across all pairs of the nine dimensions. The number of times a dimension is chosen as more relevant is the weighting of that dimension scale for a given task for that operator. The procedure permits a weighting of zero for dimensions that are judged as not relevant to workload for that task (Hart, Battiste & Lester, 1984). A workload score from 0 to 100 is obtained by multiplying the weight by the dimension scale score, summing across scales and dividing by the total weights (36 paired comparisons). The weighting procedure implicitly assumes that the individual dimensions have ratio scale properties. The weighting procedure has been found to reduce between-subject variability by up to 50% compared to unidimensional overall workload rating (Hart et al., 1984; Miller & Hart 1984).

The NASA Task Load Index (TLX) was derived from the NASA-Bipolar scales and uses a similar weighting procedure. It may be considered a shorter and more refined version. The NASA-TLX uses six dimensions, thereby considerably reducing the number of paired comparisons from 36 to 15. Aspects of task, behavior, and the operator are all included in the TLX. The first three dimensions can be considered as characteristics of the task; the next two can be considered as behavioral characteristics; and the final scale is related to the operator's individual characteristics. The six dimensions are:

- mental demand,
- physical demand,
- temporal demand,
- performance,
- effort, and
- frustration.

The descriptions of these dimensions are shown in Table 5-2. Twenty-step bipolar scales are used as the means to obtain ratings for these dimensions, as shown in Figure 5-6. Several factors were considered in choosing which dimensions to include in the TLX. Criteria such as dimension sensitivity, independence from other dimensions, and subjective importance to individual concepts of workload were considered. For ease of implementation (both in the weighting procedure and the actual rating of scales), no more than six dimensions were desired. A thorough discussion of the development of the NASA-TLX is presented in Hart and Staveland (1987).

Both the NASA-TLX and Bipolar scales have been used in laboratory and operational environments. These applications are characterized in the following descriptions. TLX was used in studies of pilot workload in helicopters (Shively, Battiste, Matsumoto, Pepiton, Bortolussi, & Hart, 1987). Four NASA test pilots flew an SH-3G helicopter on two different mission scenarios for a total of eight flights. Subjective and physiological data were collected during the flight. The TLX rating scales were administered at the end of each flight segment. During the rating, the pilot transferred control of the helicopter to the safety pilot. After rating completion, control was returned to the pilot. If control transfer to the safety pilot could not be done without excessive disruption, the pilot rating for that segment would be delayed until after the next flight segment. Pilots were never required to rate more than two consecutive segments at one time and each flight segment contained a major flight task such as hover, terrain following, or landing.

Table 5-2. NASA TLX rating scale description (NASA-Ames Research Center, 1986).

Title	Endpoints	Description
Mental Demand	Very Low/ Very High	How mentally demanding was the task.
Physical Demand	Very Low/ Very High	How physically demanding was the task.
Temporal Demand	Very Low/ Very High	How hurried or rushed was the pace of the task.
Performance	Perfect/Failure	How successful were you in accomplishing what you were asked to do.
Effort	Very Low/ Very High	How hard did you have to work to accomplish your level of performance.
Frustration	Very Low/ Very High	How insecure, discouraged, irritated, and annoyed were you.

Results indicate that TLX significantly discriminated between flight segments in both scenarios - subjective ratings and available performance measures were compared and appeared to have a relationship where a lower workload rating corresponded to better performance. Statistical analyses were not performed due to the limited amount of performance data available. However, the TLX measures appeared to be sensitive to both flight segment differences as well as performance measures.

Other applications include the use of the Bipolar scales in a laboratory study where short-term memory load, tracking task difficulty, and time-on-task were the manipulated variables (Biferno, 1985). Subjective ratings were found to correlate positively with certain physiological measures of workload. Ratings of fatigue and workload were significantly correlated for 80% of the subjects.

Bortolussi, Hart and Shively (1987) found that the Bipolar scales differentiated significantly between low and high levels of scenario difficulty in a motion-based simulator when 21 flight-related activities were added in the high difficulty scenario. These results replicate results obtained in a similar study (Bortolussi, Kantowitz, & Hart, 1986), supporting the reliability of the subjective ratings in different experiments using the same tasks but different subjects.

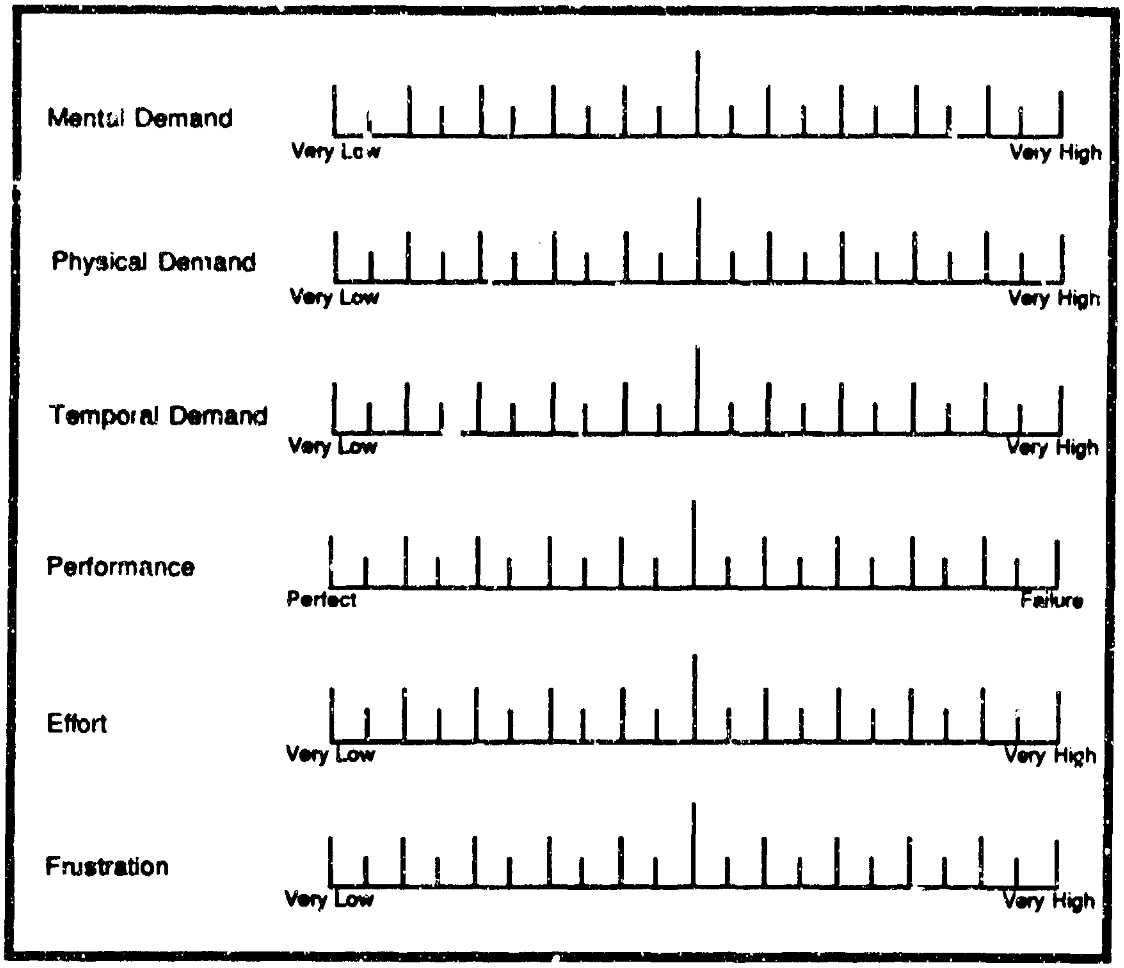


Figure 5-6. The NASA Task Loading Index (TLX) rating scales (NASA-Ames Research Center, 1986).

Vidulich and Pandit (1986) found the Bipolar scales to be sensitive to the effects of training on subjective workload ratings when the training produced lower cognitive load through development of automaticity in a category search task.

Several comparative studies have used one of the NASA scales as well as other OWL subjective techniques. The NASA scales have had high correlations with other subjective measures. Haworth, Bivens and Shively (1986) used the NASA-Bipolar scales in assessment of single pilot workload for helicopter nap-of-the-earth (NOE) missions. The correlation of NASA-Bipolar with Cooper-Harper was 0.79 and 0.67 with SWAT. In a study by Tsang and Johnson (1987), TLX and a unidimensional overall workload scale followed very similar trends. Vidulich and Tsang (1987) found similar correlations among TLX, an overall workload scale, and the Analytic Hierarchy Process.

Vidulich and Tsang (1985a, 1985b, 1986) compared the subjective measures obtained from NASA-Bipolar and SWAT for both tracking and spatial transformation tasks. Both techniques were found to be sensitive to various levels of task demands and generally provided similar results. A comparison of the two techniques shows that the NASA-Bipolar scales result in less between-subject variability but use more dimensions of workload (although, at the time, it was unclear whether all nine dimensions added information). NASA-Bipolar required less time in the weighting procedure compared to SWAT's scale development procedure, but more in actual ratings because of the nine scales as compared to three dimensions of SWAT. A similar comparison between NASA-TLX and SWAT has not yet been reported in the literature.

Of the two NASA scales, the TLX scale is the version that is recommended by NASA. Information for administration of TLX is contained in *Collecting NASA Workload Ratings: A Paper-and Pencil Package* (NASA-Ames Research Center, 1986). Contained in this package are copies of the six rating scales, the fifteen paired-comparisons, sources of workload tally sheets and instructions on the procedures to follow to obtain individually weighted workload scores. A computerized version is also available which provides software that will display the rating scales, tally the sources of workload, and provide the weighted scores. Being newer, not as much research has been reported for the TLX version as for the full Bipolar version. Further research and application examples will provide additional information with which to characterize the TLX scale fully.

As with other multidimensional scales, not only can an overall workload score be obtained, but the individual scales could be used to diagnose what aspects of workload were particularly relevant for a specific task. The ability to identify what task, behavior, or operator characteristic was judged to have the greatest impact on the perception of workload would provide an additional diagnostic tool to assess system design alternatives.

Summary. Both the NASA-Bipolar and TLX scales have been proven to be valid, reliable and sensitive techniques for OWL assessment. The scales have been used in laboratory and applied settings. The multidimensional nature of workload and the relevance of various workload dimensions to individual assessment of workload are both accounted for in the individual weighting procedure and six dimensions used in TLX. TLX was derived from the Bipolar scales and is the technique currently recommended by NASA. Certainly, the approach used by both scales is useful. The TLX is more practical for operational applications because it is shorter and takes less time to complete. TLX is only beginning to be characterized although the validity of its underlying approach is supported by its predecessor's (Bipolar scales) research base.

Psychometric Techniques

Among the rating scale techniques available are those that are based on classic psychometric scaling methodologies. Psychologists have used these methods as a means of quantitatively measuring psychological attributes. Workload might be considered to be such an attribute. Among the best known are magnitude estimation, paired comparisons, and equal appearing intervals.

Magnitude estimation. Magnitude estimation is a psychophysical method that requires a subject to make direct numerical assignments to the magnitude of some sensory experience. It is one of the most frequently used psychophysical scaling methods (Gescheider, 1985). There are two main procedures for obtaining magnitude estimation (Stevens, 1958; 1975). In the first method, a subject is presented with a standard stimulus and is told this experience represents a certain numerical value (called a modulus). The subject is then asked to make judgments relative to the modulus. For example, if the modulus is assigned 10 and the experience is judged to be twice as great as the one created by the standard stimulus, the subject would say 20. In the second method, the modulus is not defined by the experimenter and the subject is asked, in essence, to establish his own modulus.

Some research has used magnitude estimation in workload assessment (e.g., Borg, 1978; Helm & Heimstra, 1981). High correlations have been reported between subjective estimates of workload and task difficulty (e.g., Helm and Heimstra used information load (bits/sec) as the measure of task difficulty). Masline (1986) found equal sensitivity among magnitude estimation, equal appearing intervals, and conjoint scaling as used in SWAT. Gopher and Braune (1984) describe the use of magnitude estimation scaling for workload assessment in 21 experimental conditions. A single-axis tracking task was used as the modulus and given a value of 100. After each trial, subjects were asked to estimate the load or demand of other tasks. Gopher and Braune found that subjects did not have any difficulty in assigning numbers despite the wide variety of tasks. They also constructed a power function and used it to predict the loads of dual tasks from single task scores. They found a high correlation between resource requirements (derived from subjective scores) and an index of task difficulty, but low correlations with reaction time performance measures.

The magnitude estimation method was also used by Kramer, Sirevaag and Braune (1987) to collect subjective ratings of OWL in a single-engine, fixed-base simulator. A five minute straight and level flight path segment was used as the modulus and assigned a value of 100. These researchers found that the subjective ratings corresponded well to flight task performance, as measured by flight heading and altitude deviations, reaction time, and accuracy of the auditory secondary probe task. Both subjective ratings and performance measures differentiated between easy and difficult flights and between flight segments. However, the pilots' workload estimates indicate that holding patterns, takeoffs, and landings were equally difficult, while performance measures indicated holding patterns and straight-level-flight were done better than takeoffs and landings.

One of the ways of using magnitude estimation is to have a standard reference task (a modulus) against which relative judgments of workload are made. Rather than allowing subjects to make relative judgments against their own internal reference developed from past experience, Hart and Staveland (1987) suggest that a standard reference task may reduce between-subject variability. They suggest that reference tasks may assist in providing a stable judgment set from which to make estimates of subjective workload. They also suggest that the reference task should share elements with the experimental tasks to be performed, because the workload of different tasks may be caused by different task dimensions. The reference task should provide the opportunity to make comparable judgments.

O'Donnell and Eggemeier (1986) review workload assessment that has used magnitude estimation and they conclude that the data support the estimates obtained from magnitude estimation techniques. They do caution, however, that the use of magnitude estimation may have practical limitations. For example, subjects may not be able to retain and use the same modulus over time. In addition, counterbalanced presentation of stimuli, normally used in laboratory magnitude estimation experiments may not be possible. O'Donnell and Eggemeier (1986) suggest that the impact of these potential problems should be identified before magnitude estimation techniques are used in operational environments.

Paired comparisons. Other psychometric techniques that might be used for workload assessment include paired comparisons (also called Thurstone scaling techniques). In the paired comparison technique, subjects choose one of a pair of stimuli which has more of the characteristic being judged. The number of comparisons made is n(n-1)/2, where n is the number of stimuli. Therefore, the number of comparisons can become quite large as the number of stimuli increases. Five stimuli would require 10 comparisons; eight stimuli would require 28 comparisons; and ten would require 45 comparisons. Scales are derived from the number of times a stimulus is judged to have more of the relevant characteristic than the other stimuli.

Equal-appearing intervals. The technique of equal-appearing intervals has the subject assign the stimulus to one of several categories depending on how much of a characteristic the stimulus is judged to possess. Eleven categories are often used. The subjects are also instructed to keep the distance between any two categories equal to the distance between any other two. Hicks and Wierwille (1979) applied this technique in a study of workload in an automobile simulator. Results indicated significant differences between all task difficulty levels which indicate the method is sensitive to workload variations. Masline (1986) found the sensitivity of the equal-interval technique to be equivalent to magnitude estimation and to SWAT. He concluded that the equal-interval scaling was the easiest of the three to administer. Masline cautions that there is a strong tendency for operators to assign stimuli so that all categories are used about equally often, which may bias ratings.

Summary. The psychometric techniques of magnitude estimation, paired comparisons, and equal-appearing intervals have been used as workload assessment techniques. In general, studies have indicated sensitivity of the methods to varying task difficulty levels. The psychometric techniques appear to offer viable alternatives for subjective workload assessment although reservations about the use of these techniques in operational environments were expressed by O'Donnell and Eggemeier (1986). More information on the development and procedures for using these techniques is required before they are fully recommended for Army applications. For further information, the reader should consult texts on psychophysical methods (e.g., Edwards, 1957; Stevens, 1975; Gescheider, 1985) as well as reviews of these techniques as applied to workload assessment (e.g., O'Donnell & Eggemeier, 1986).

Subjective Workload Assessment Technique (SWAT)

The Subjective Workload Assessment Technique (SWAT) is a subjective rating technique developed by the U.S. Air Force Armstrong Aeromedical Research Laboratory (AAMRL) at Wright-Patterson Air Force Base. It uses conjoint measurement and scaling techniques (Krantz & Tversky, 1971; Nygren, 1982) to develop a rating scale with interval properties. SWAT uses the three dimensions of time load, mental effort load, and psychological stress load to assess workload. These were adapted from the workload definition developed by Sheridan and Simpson (1979). For each of the three dimensions, there are three levels which are operationally defined. These are shown in Table 5-3. *Time load* refers to the relative amount of time available to the operator (AAMRL, 1987) and the percentage of time an operator is busy (Eggemeier, McGhee, & Reid, 1983), and includes elements such as overlap of tasks and task interruption. *Mental effort* refers to the amount of attention or concentration directed toward the task, independent of time considerations. *Psychological stress* is the degree to which confusion, frustration and/or anxiety is present and adds to the subjective workload of the operator. Factors that may increase stress and elevate distraction from the task include personal factors such as motivation, fear, fatigue or environmental factors such as temperature, noise or vibration (AAMRL, 1987).

There are two distinct steps in the use of SWAT. The first is called scale development. Twenty-seven cards contain all possible combinations of the three levels of each of the three dimensions. The cards are sorted by the individual operators into the rank order that reflects their perception of increasing workload. The SWAT User's Guide (AAMRL, 1987) suggests that the 27 cards be first sorted into three piles of nine each, and then each pile ordered 1 through 9 representing lowest to highest workload. The order of the sorted cards are then processed via conjoint scaling procedures to develop a scale with interval

Table 5-3. Operational definitions of the three SWAT dimensions (AAMRL, 1987).

LEVELS	DIMENSION
	I. TIME LOAD
1	Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.
2	Occasionally have spare time. Interruptions or overlap among activities occur frequently.
3	Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.
	II. MENTAL EFFORT
1	Very little conscious mental effort or concentration required. Activity is almost automatic requiring little or no attention.
2	Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.
3	Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.
	III. PSYCHOLOGICAL STRESS
1	Little confusion, frustration or anxiety exists and can be easily accommodated.
2	Moderate stress due to confusion, frustration or anxiety. Noticeably adds to workload. Significant compensation is required to maintain adequate performance.
3	High to very intense stress due to confusion, frustration or anxiety. High to extreme determination and self-control required.

properties. The developed numerical scale runs from 0 to 100, with 0 signifying no workload, or the lowest ranked condition on each of the three dimensions (usually 1,1,1), and 100 corresponding to the maximum workload, or the highest ranked on each of the three dimensions (usually 3,3,3). Other combinations of ratings on the three dimensions (e.g., 2,3,2) would be assigned a corresponding scale number (e.g., 75). The scale value corresponding to each combination of rating will be different for each individual dependent on the way the cards are sorted. An illustration of the mapping from a three dimensional to a unidimensional scale is shown in Figure 5-7.

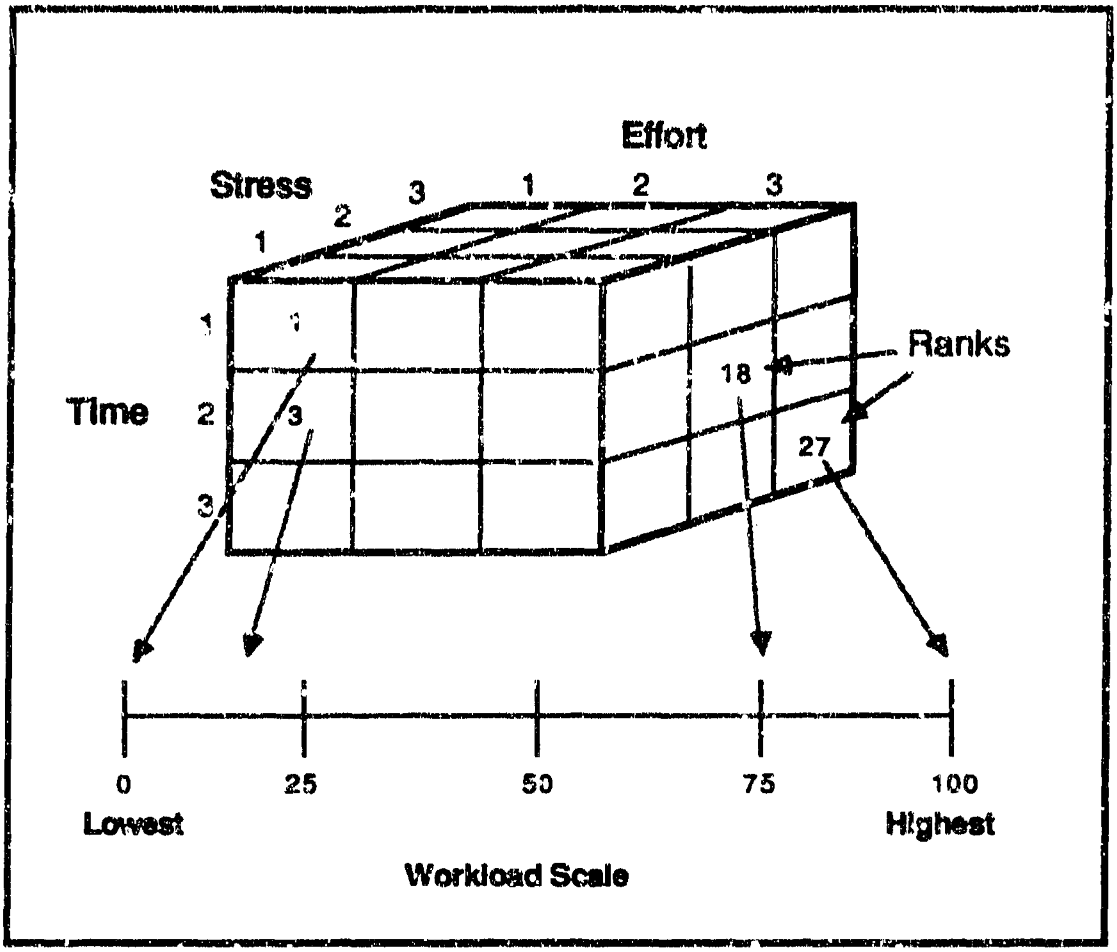


Figure 5-7. Subjective Workload Assessment Technique (SWAT) uses conjoint analysis to change each individual's ranks to a unique interval scale. In this individual example, the rank of 1, given to the combination 1,1,1, is reflected as the lowest (0) value on the Workload Scale. The rank of 27, given to the combination 3,3,3, is reflected as the highest (100) value on the Workload Scale. Intermediate rank values of 3 and 18 given to the combinations 2,1,1 and 2,3,2 respectively, are reflected as intermediate workload values (i.e., 20 and 75, respectively) dependent on the individual workload scale developed. (The illustration is adapted from Gidcumb, 1985.)

The second step to SWAT is the event scoring, that is, the actual rating of workload for a given task or mission segment. For the defined task or segment, the operator is asked to assign a level (1, 2 or 3) to each of the three dimensions of time load, mental effort, and psychological stress. It has been found that the order in which the three dimensions are presented does not affect the rankings (Acton & Colle, 1984), but it is suggested that the order in which they are ranked be kept constant to reduce confusion (AAMRL, 1987). This rating is converted to one of the 27 numerical scores (described above) between 0 and 100 which are computed during scale development.

Since the initial development of SWAT, there have been many refinements, suggestions for implementation, analysis, and interpretation. One issue of concern is the difference between individual and group scale development. (The group scale is constructed using group mean rankings.) In an early SWAT study, there were high coefficients of concordance for the rankings of four different groups of operators ranging between .76 and .82 (Reid, Shingledecker, & Eggemeier, 1981). Because of the high level of agreement, a group scale was developed for each different experiment. With group scales, the idiosyncrasies of individual sorts tend to average out. Conversely, the group scale may also hide some of the individual differences in the perception of workload. An alternative approach has been developed that permits scale development for homogeneous subgroupings of individuals called *prototypes* (Reid, Eggemeier, & Nygren, 1982). The prototypes are based upon which one of the three dimensions is the overriding factor in their rankings. For example, if time load is considered as most important, the rankings may reflect a certain level of time held constant while the other two dimensions are varied across the full range of possibilities. The SWAT User's Guide (AAMRL, 1987) discusses specific procedures and approaches to use in determining how individuals should be grouped together into time, effort or stress prototypes. The prototype approach offers an increased sensitivity to individual differences as compared with the group approach.

Time, effort and stress may be individually examined as workload components -- whether individual, group, or prototype sorts are used. How the particular dimensions are rated may be useful in determining the specific design features that may be contributing to the workload perception. If the time load dimension is judged to be very high while the other two are not scored as high, for example, this might suggest that a design element in the time domain (e.g., data presentation rate or required response time) is the most important consideration for workload in that task or mission segment (Eggemeier, McGhee, & Reid, 1983).

SWAT meets many of the practical considerations for use of workload assessment techniques. As with other subjective techniques, such considerations include ease of implementation, high face validity, operator acceptance, relative freedom from interference with the primary task (i.e., intrusiveness), scorability (i.e., the degree to which it can be quantified), repeatability and quickness of administration (Crabtree, Bateman, & Acton, 1984; Courtright & Kuperman, 1984).

There have been numerous studies of SWAT as a workload measurement technique in both laboratory and applied settings. Laboratory studies have shown that SWAT is sensitive to differences in task demands in critical tracking and simulated aircrew radio communication tasks (Reid et al., 1981); continuous recall tasks (Potter & Acton, 1985); a spatial memory task (Eggemeier & Stadler, 1984); a short-term memory task (Eggemeier, Crabtree, Zingg, Reid, & Shingledecker, 1982); simulated air-to-air combat (Reid, Eggemeier, & Shingledecker, 1983, cited in Eggemeier, McGhee, & Reid, 1983); and a probability monitoring task (Notestine, 1984).

Most of the applied studies have used SWAT in aviation applications. This is certainly not surprising given the Air Force roots of SWAT and the traditional concern with pilot workload. Skelly and Purvis (1985) used SWAT in an investigation of a B-52 wartime mission simulation. Haworth, Bivens and Shively (1986) used SWAT in a single pilot helicopter, nap-of-the-earth flight simulation. Gidcumb (1985) reports the use of SWAT in several Air Force applications. Courtright and Kuperman (1985) discuss the use of SWAT in Air Force test and evaluation environments and found the technique understandable and accepted by both testers and subjects. Schick and Hann (1987) used a German-language version of SWAT to assess workload in a moving-base cockpit simulator. They report that SWAT was sensitive to varied task difficulty.

However, application of SWAT has not been limited to aviation environments. Crabtree, Bateman and Acton (1984) used SWAT in an examination of over 20 command, control and communication (C^3) tasks (the tasks are not described in detail). SWAT ratings were also obtained in a study of the effects of experience level on the performance of nuclear power control room crews (Beare & Dorris, 1984).

The reliability of the SWAT card sorts has been typically found to be high: the correlation ranged from .77 to 1.00 for four pilots for pre- and post-test card sorts (Gidcumb, 1985). Subjects have produced card sorts as far apart as a year, and over eighty percent of the subjects produced sorts that correlated .90 or above (AAMRL, 1987). These correlations suggest stable workload judgments will be made across time.

With SWAT, as with other subjective techniques, there is a question regarding the effects of delays between the workload experience and the rating. Some research has specifically looked at this question and concluded that although there were some changes in ratings, short delays of 15-30 minutes do not affect the overall mean ratings (Eggemeier, Crabtree, & LaPointe, 1983). This may have been due to a counterbalancing effect where some subjects increased rating and others decreased ratings relative to the baseline. Eggemeier, Melville and Crabtree (1984) found that neither 14-minute delays nor intervening tasks affected subjective workload ratings. However, delayed ratings should not be expected to be exactly the same as ratings given immediately after performance. This is particularly important if the absolute value is desired, but not as important if relative values are desired for comparison between two alternative task or equipment configurations. Additionally, it was found that the most difficult intervening task produced the most discrepant ratings. This finding is troubling because of the analogy that can be drawn to applied studies. Often the reason for operators not providing ratings when asked is that they are too busy with a difficult, high workload task. At the next occasion for rating, a difficult intervening task will have occurred, therefore, the research suggests that the rating will be lower than it would have been without the delay caused by the difficult task (Eggemeier, Melville, & Crabtree, 1984).

SWAT appears to be a valid measure of some aspects of workload, particularly those associated with the operationally defined dimensions of time load, mental effort and psychological stress. SWAT has been found to give similar results to other subjective methods such as NASA bipolar ratings (Vidulich & Tsang, 1986; Haworth et al., 1986); the Modified Cooper-Harper scale (Warr et al., 1986); magnitude estimation and equal-interval scales (Masline, 1986); as well as to compare favorably with various physiological measures (Albery, Repperger, Reid, Goodyear, & Roe, 1987).

Several practical observations and suggestions were made by Gidcumb (1985) in a report that used SWAT as a workload measure in several Air Force applications. He concluded that "SWAT appears to be an accurate measure of the workload experienced by the aircrew participating in the tests surveyed" (p. V-1). However, several suggestions were made to improve the use of SWAT in applied settings. During the introductory briefing to SWAT, more emphasis should be placed on what will be expected of the operators. There were observations that some of the operators approached the card sorting task very casually, as evidenced by cursory card sorts. Gidcumb suggests that operators be fully introduced to the benefits of SWAT to them personally and the importance of the card sort to the entire procedure. The motivation of the operators is a critical element in the success of the card sort.

SWAT administrators agreed that the operators should be thoroughly familiar with the rating procedures, and after six to ten SWAT ratings aircrew felt confident that the ratings were reflecting their workload perceptions. After 15 ratings, the aircrew reported the ratings interfered little with their other duties. Practice with the rating procedure and the operational definitions of the dimension levels is very important in obtaining accurate workload measures. Without adequate practice, a learning effect may distort the ratings and both relative comparisons and absolute measures of workload will have only limited value.

There were other comments dealing with the gathering of ratings. Some pilots refused to consider real-time ratings because of their concern that it would impose an additional task. An alternative procedure was used where the pilots would review mission videotapes for post-flight ratings. Another way to handle missing ratings would be to assign the highest rating (3,3,3) for real-time segments that were missed (AAMRL, 1987). The operators also had trouble deciding what to rate segments that were impossible to perform or were performed incorrectly. The suggestion was made that the SWAT administrator needs to be explicit about what to rate and what kind of ratings should be assigned impossible or differently performed tasks.

Schick and Hann (1987) suggest that SWAT data collection be planned carefully so that obtaining ratings does not interfere with flight duties. Therefore, event-related data collection (as has been used in most studies) appears to be a better alternative than data collection at fixed time intervals. Interestingly, this is different from Gidcumb's (1985) recommendation that further research should be done on time-based rather than task-based rating segments.

Another observation was that in these applications (and, it can be inferred, most operational applications), only a small number of operators and data gathering missions are available, therefore sample sizes are small. Parametric statistical analyses may be inappropriate and other descriptive or comparison techniques may be more appropriate in such cases.

Other issues involve the expansion of the current rating scheme from the three current levels to, perhaps, five. Although this might provide greater rating sensitivity and avoid floor or ceiling effects (as suggested by Potter & Acton, 1985), the card sort (as currently administered) with five levels might become unmanageable for subjects. However, finding some approach to increase the number of levels may yield benefits. The use of partial sorts, consisting of a subset of the original number of combinations, may be a possible method, although this has not as yet been thoroughly developed (Nygren, 1985).

There is also a question of the ability needed to perform the card sort procedure. It is recognized that the card sort is the key to successful use of SWAT and that motivation does play a role in how carefully the cards are sorted. The cards contain combinations of verbal descriptions and there is some anecdotal evidence to suggest that individuals with *low verbal skills* may have difficulty in the sorting task. A possible solution would be the use of graphical representations. This is an area for further investigation -- empirical data are needed to examine this issue to see if it is a problem. Solutions will be proposed and investigated if this is proven to be a problem (G. B. Reid, personal communication, July 9, 1987).

Several other concerns have been raised in addition to the potential problems associated with the card sorting procedure. It has been suggested (Derrick, 1983; Hart, 1986a) that three factors may not be enough to adequately characterize workload. The three factors, it has been suggested, may not be orthogonal (Boyd, 1983). Hart (1986a) discusses that the assumption that people can accurately distinguish between the 27 combinations may not be true. Boyd (1983) suggests that there might be high interrater reliabilities at the extremes, but the intermediate ratings may be less reliable. A further concern is that scales with fewer than six or seven divisions may have response nonlinearities near the endpoints (Hart & Staveland, 1987).

Summary. The Subjective Workload Assessment Technique (SWAT) uses conjoint analysis to obtain a workload rating scale with interval properties. SWAT uses the three dimensions of perceived time load, mental effort, and psychological stress to assess OWL. Both scale development and an event scoring procedures are used. These provide individual rank order of dimension and ratings on the three dimensions for a given task or task segment. SWAT has been shown to be both valid and reliable as a measure of workload. SWAT has been used in both laboratory and applied settings and found to be sensitive to a variety of task demands. Because of the multidimensional nature of SWAT, it is possible to use the individual dimension scales as diagnostic OWL tools. Care must be taken in the card sort and event-scoring implementation to obtain accurate workload measures. SWAT appears to be a useful technique for subjective workload assessment in Army applications.

Other Subjective Rating Scales

There are other rating scales that have been developed for workload assessment. Often scales are created for specific studies. This has led Shingledecker (1983, as cited in Potter, 1986) to suggest that there may be almost as many scales and checklists as there are studies that use subjective assessment techniques. Of these many subjective techniques, a few are presented here as examples of other types of rating scales that have been developed and used in workload assessment applications.

The Pilot Subjective Evaluation. The Pilot Subjective Evaluation (PSE) process was developed by Boeing for use in the workload evaluation of the Boeing 767 (Fadden, 1982; Ruggiero & Fadden, 1987). The PSE is shown in Figure 5-8. It includes both seven-point rating scales and an accompanying questionnaire. A validation study of the PSE is reported although details are not given in either of these two papers. The particularly interesting aspect of these scales is the use of a reference airplane (chosen by the pilot) for a comparative evaluation of workload. Basically, the pilots rated whether operation of the 767 was more, the same, or less demanding than than the reference aircraft in terms of mental effort, physical difficulty, and time required. Ratings of greater workload indicate areas for design improvements. An interview, held at the end of each day, provided the opportunity to gather more information on the items rated worse than the reference airplane workload.

The Dynamic Workload Scale. The Dynamic Workload Scale is another rating scale developed for an aircraft certification program and has been used by Airbus Industrie (Speyer, Fort, Fouillot, & Blomberg, 1987). As seen in Table 5-4, the scale is a seven-point scale. The technique includes workload assessment by both the pilot and an observer-pilot. The scale is administered without defining workload, allowing the pilot and observer to be guided by their own interpretation of workload. However, the criteria for the raters to consider are reserve capacity, interruptions and effort or stress. The observer makes a rating whenever the workload has changed since the last rating or when five or more minutes have passed. A cue is then given to the pilot to make a rating. The primary analyses of these data were cumulative rating distributions. Concordance between pilot and observer ratings were also examined and appeared high. Ratings are also plotted along a timeline. Speyer et al. (1987) report a shift in the median of the distribution of ratings as workload increased, implying a sensitive measure, although no further details are available.

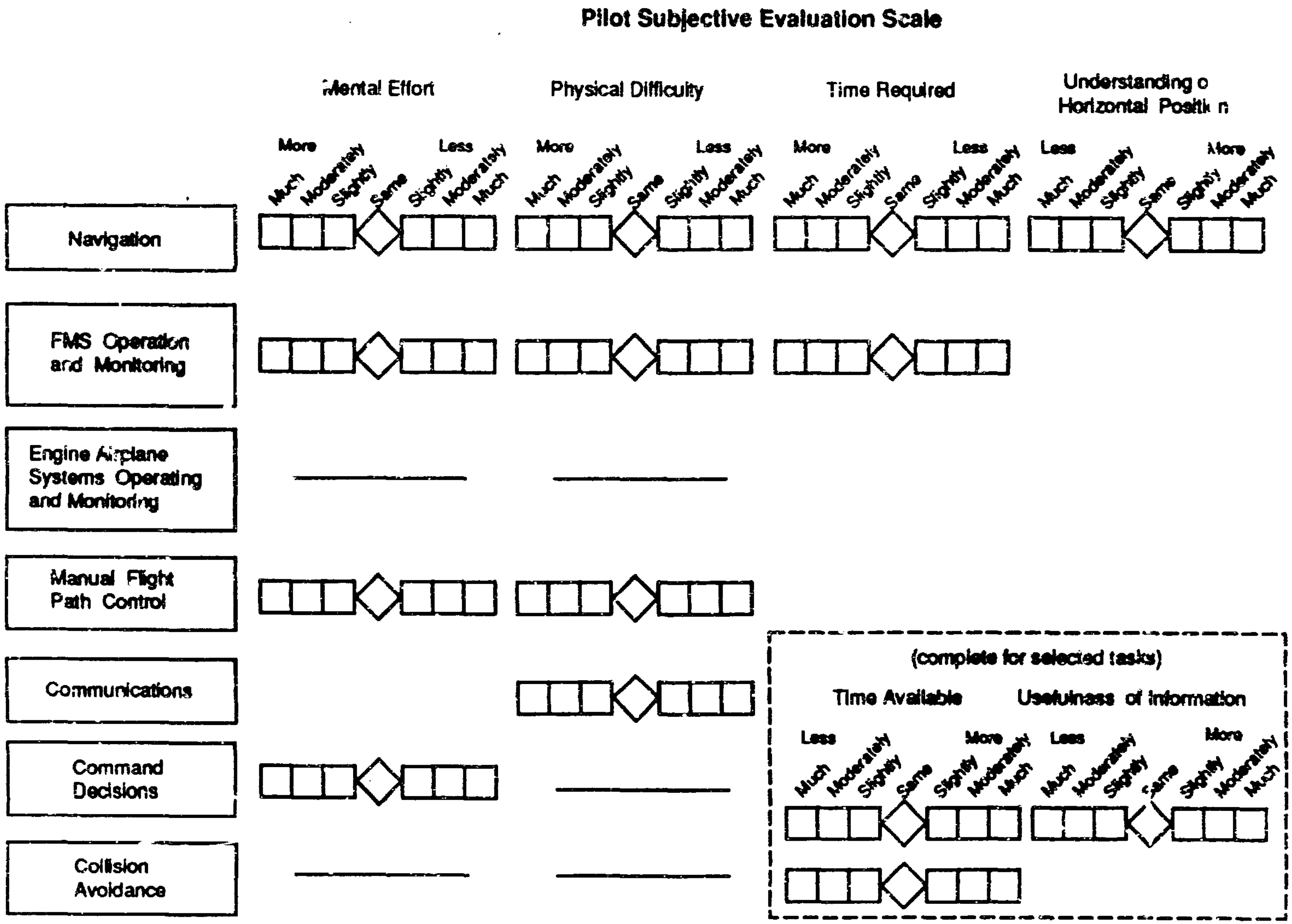

Pilot Subjective Evaluation Scale
Mental Effort
Physical Difficulty
Time Required
Understanding o Horizontal Positi n
More
Less
Much
Moderately
Slightly
Same
Slightly
Moderately
Much
Navigation
FMS Operation and Monitoring
Engine Airplane Systems Operating and Monitoring
Manual Flight Path Control
Communications
Command Decisions
Collision Avoidance
(complete for selected tasks)
Time Available
Usefulness of information

Figure 5-8. The Pilot Subjective Evaluation Scale Developed by Boeing.

Table 5-4. The Dynamic Workload scale used by Airbus Industrie for aircraft certification (Speyer et al., 1987).

Workload Assessment	Criteria			Appreciation
	Reserve Capacity	Interruptions	Effort or Stress	
Light 2	Ample	————	————	Very Acceptable
Moderate 3	Adequate	Some	————	Well Acceptable
Fair 4	Sufficient	Recurring	Not Undue	Acceptable
High 5	Reduced	Repetitive	Marked	High but Acceptable
Heavy 6	Little	Frequent	Significant	Just Acceptable
Extreme 7	None	Continuous	Acute	Not Acceptable Continuously
Supreme 8	Impairment	Impairment	Impairment	Not Acceptable Instantaneously

Analytic Hierarchy Process. A scaling procedure which uses paired comparisons is based on the Analytic Hierarchy Process (AHP) developed by Saaty (Lidderdale, 1987). The procedure was aimed at obtaining relative estimates of workload after flights. All possible pairs of tasks or task segments are presented to the operator (in this case, the pilot). If one of the pair is judged to have higher workload, the operator is asked to judge by how much on a scale from 1 to 5:

1 = equal workload

2 = slightly higher workload

3 = moderately higher workload

4 = very much higher workload

5 = extremely high relative workload.

Through mathematical procedures (Lidderdale, 1987; Lidderdale & King, 1985; Saaty, 1980), the ratings can be used to obtain relative judgments of mission element workload. Visual inspection of graphs of workload assessments of the same mission elements obtained by the Bedford scale and the AHP method show similar results and a rank order analysis gave high correlations (Lidderdale, 1987).

Vidulich and Tsang (1987) classify the AHP as a relative judgment method as opposed to the absolute judgments of workload that are obtained with NASA-TLX or a unidimensional workload scale. All three OWL scales were used in a single-axis compensatory tracking task with control order determining the level of task difficulty and visual or auditory presentation. All three OWL measures exhibited close agreement in discriminating the task variables, although the AHP showed the greatest validity (as measured by correspondence to performance) and reliability (as measured by test-retest correlations). However, topics of concern include how well relative judgments could be made across more varied tasks, as well as the possibility of subjects forgetting details or creating their own hypotheses about task relationships. Vidulich and Tsang suggest that further research with the AHP should be pursued.

Workload/Compensation/Interference/Technical Effectiveness. The Mission Operability Assessment Technique (MOAT) is another technique that uses conjoint scaling methods (Donnell, 1979; Helm & Donnell, 1979). The MOAT process was designed to evaluate overall system operability, specifically in aviation environments. As part of the MOAT process, the Workload/ Compensation/ Interference/ Technical Effectiveness (WCI/TE) matrix and rating scale was developed. The WCI/TE is a 4 X 4 matrix which describes technical effectiveness of the system (4 levels) and pilot workload, compensation and interference (4 levels) and is shown in Figure 5-9. As in all conjoint scaling techniques, pilots first rank order the 16 matrix elements and then specific tasks are rated. The task rating can then be transformed to an interval value from 0 to 100.

Some data on the sensitivity of the WCI/TE are available from work done by Wierwille and Connor (1983). The study used psychomotor tasks in a moving-based flight task simulator. The WCI/TE scale was found to significantly differentiate between three levels of task difficulty. Wierwille et al. (1985) also report the WCI/TE to be generally sensitive to psychomotor, perceptual, and mediational tasks (the WCI/TE was not tested with communication tasks). O'Donnell and Eggemeier (1986) suggest that MOAT was specifically intended for piloting tasks and was not intended as a direct measure of workload.

Summary. These techniques represent several additional subjective workload assessment tools. The PSE and the Dynamic Workload Scale were developed specifically for civilian aircraft certification and provide interesting examples of applied techniques. The WCI/TE scale has been found to be a sensitive workload measure (Wierwille et al., 1985), but currently appears to be of interest only as the conjoint scaling predecessor to SWAT.

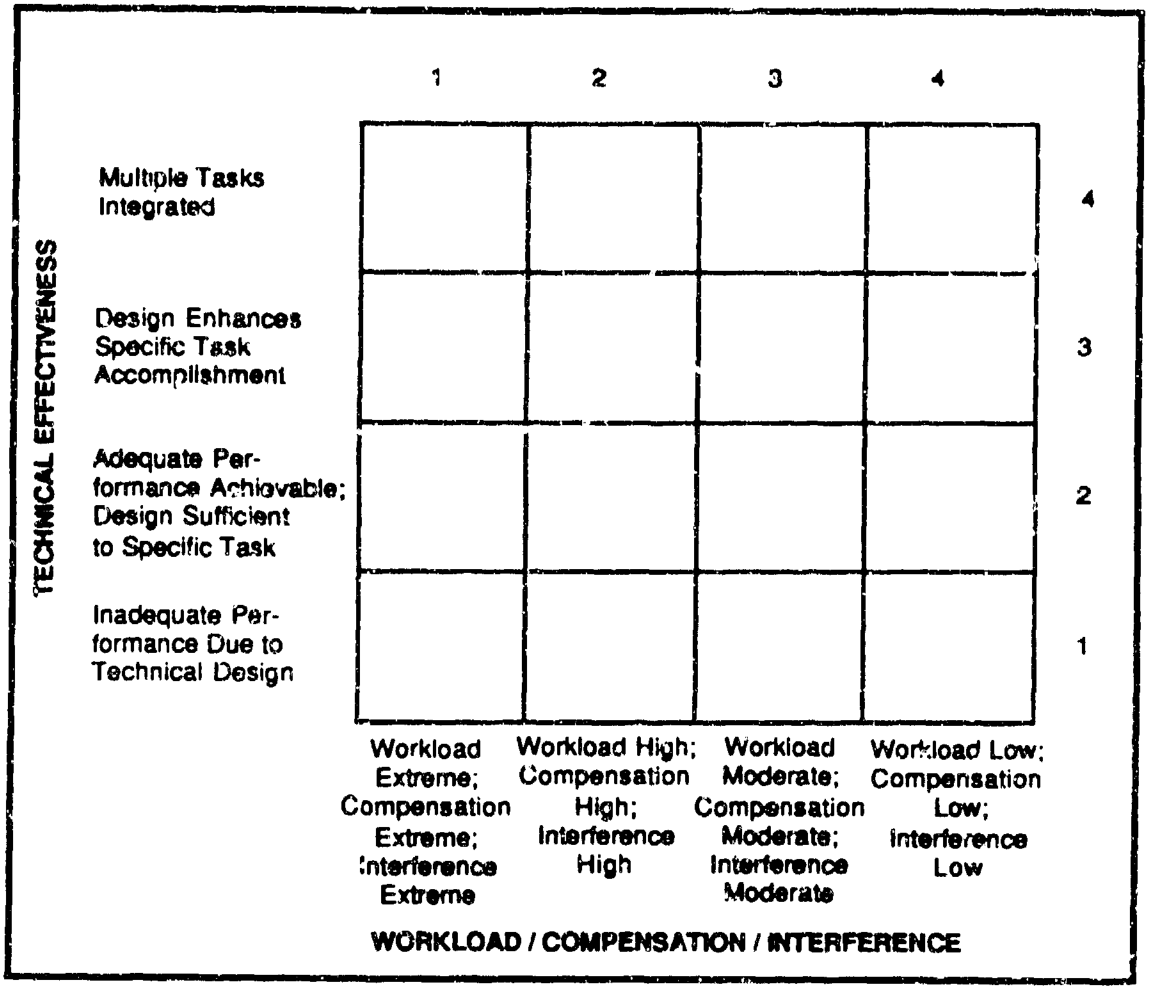

1
2
3
4
TECHNICAL EFFECTIVENESS
Multiple Tasks Integrated
4
Design Enhances Specific Task Accomplishment
3
Adequate Performance Achievable; Design Sufficient to Specific Task
2
Inadequate Performance Due to Technical Design
1
Workload Extreme; Compensation Extreme; Interference Extreme
Workload High; Compensation High; Interference High
Workload Moderate; Compensation Moderate; Interference Moderate
Workload Low; Compensation Low; Interference Low
WORKLOAD / COMPENSATION / INTERFERENCE

Figure 5-9. The WCI/TE scale matrix (Donnell, 1979; Helm & Donnell, 1979).

The AHP is a technique that has recently been used for workload assessment. Lidderdale (1987) found it useful in an applied setting, while Vidulich and Tsang (1987) found it more reliable and valid than two other scales in a single and dual-task laboratory experiment. Sufficient information is not yet available to make judgments on the AHP for Army OWL assessment. Further research is needed.

Comparisons Among Rating Scales

The results of comparisons among different rating scales have been briefly described in previous sections. There have been several additional studies that have directly compared more than one subjective measure of workload. Some have used multiple measures as a battery of workload assessment tools -- others have performed research intended as comparisons and validation studies of the various techniques. Table 5-5 presents a matrix of the subjective OWL techniques that have been discussed in the previous sections. Within the matrix, published research that has used more than one subjective technique is listed. Although it is believed that the primary comparative studies are listed, some research that could have been included in this table may not have been identified. However, the table is representative of the research that has been done and the gaps that still exist.

The literature altogether indicates that the techniques that have been compared correspond well. Generally, the same rank order of task difficulty are obtained by each technique. Each of the studies listed in Table 5-5 is briefly presented:

- Hart and Staveland (1987) describe the development of NASA-TLX as a refined edition of the NASA-bipolar scales. NASA-TLX was developed to reduce the number of scales (from ten to six) by selecting those dimensions that best discriminated between task variables, that provided independent information, and were associated with overall workload ratings. NASA-TLX and bipolar scales were not compared as such, but the relationship between the scales and supporting empirical data are presented in detail.

- Haworth, Bivens and Shively (1986) investigated workload in single-pilot operation in NOE helicopter missions. They used the Cooper-Harper scale, the NASA-Bipolar scales and SWAT to assess handling qualities (using CH) and workload. The correlation between the Bipolar and SWAT was .67, while CH was significantly correlated to NASA-Bipolar and SWAT measures (.75 and .79, respectively). Both subjective techniques indicated a higher average workload for one pilot as compared to two pilots.

- Lidderdale (1987) used the Bedford scale to obtain real-time OWL ratings for an advanced combat aircraft with a two-person crew during low level maneuvers. The Saaty AHP was used to obtain OWL ratings in a post-flight context. Visual inspection of graphs of workload ratings for each flight segment by both techniques show similar results. The Spearman rank order correlation between the Bedford and AHP scores

Table 5-5. Workload studies that have used more than one rating scale.

Modified CH	NASA BIPOLAR	NASA TLX	SWAT	AHP	OVERALL WORKLOAD	WCI/TE	RATING SCALE
	HAWORTH, ET AL. (1986)		HAWORTH, ET AL. (1986)			WIERWILLE AND CONNOR (1983)	COOPER HARPER
*WIERWILLE, SKIPPER AND RIEGER (1984)			WARR, ET AL. (1986)			WIERWILLE, ET AL. (1985)	Modified CH
		TSANG AND JOHNSON (1987)		LIDDERDALE (1987)	TSANG AND JOHNSON (1987)		BEDFORD
		** HART AND STAVELAND (1987)	VIDULICH AND TSANG (1985A,B; 1986) HAWORTH, ET AL. (1986)				NASA BIPOLAR
				VIDULICH AND TSANG (1987)	VIDULICH AND TSANG (1987)		NASA TLX
			MASLINE (1986)				MAGNITUDE ESTIMATION
			MASLINE (1986)				EQUAL INTERVAL
							SWAT
					VIDULICH AND TSANG (1987)		AHP
							OVERALL WORKLOAD

* WIERWILLE, SKIPPER AND RIEGER (1984) COMPARED FIVE DIFFERENT FORMS OF THE MODIFIED CH.

**THE NASA-TLX IS A MORE REFINED VERSION OF NASA-BIPOLAR. SEE HART AND STAVELAND (1987) shows significant correlation coefficients of .86 for the pilots and .85 for the navigators. The AHP obtains a relative, post-flight assessment, while the Bedford scale is considered an absolute scale. The author suggests, however, that the Bedford scale may be considered relative in that workload assessments are probably made from a baseline of all previous experience.

- Masline (1986) used SWAT, magnitude estimation, and equal-appearing interval scales to assess workload of a continuous recall task where presentation rate, number of digits and the number of positions back to recall were varied. Results indicated equal sensitivity among the three techniques. Correlations between subjective and performance measures were significant. Masline compared the three techniques with other criteria: all three appeared equivalent in terms of sensitivity, predictive capability, obtrusiveness and operator acceptance. However, SWAT appeared to have greater diagnosticity because of its multidimensional nature. The easiest technique to administer was the equal-interval scale.

- Tsang and Johnson (1987) used a battery of three subjective measures to assess workload in several manual and semi-automated tasks. The NASA-TLX scale, the Bedford scale, and a unidimensional overall workload scale were used. The NASA-TLX and overall workload scales displayed very similar trends for the different tasks. Interestingly, the operator workload ratings showed a training effect evidenced by a decrease in ratings in later sessions (i.e., over three sessions). The authors suggest these findings demonstrate the sensitivity and robustness of these measures.

  The slightly different ratings obtained from the Bedford scale were interpreted, in light of multiple-resource theory (Wickens, 1980), as supportive of the ability of the Bedford scales to assess what it claims to assess, that is, spare capacity. The authors do caution that these conclusions are based on limited data from only six subjects.

- Vidulich and Tsang (1986) (see also Vidulich & Tsang, 1985a & 1985b) used both NASA bipolar scales and SWAT ratings to assess workload in a laboratory study using tracking and spatial transformation tasks. Both techniques displayed sensitivity to the various task demands and, in general, provide similar results. Haworth et al. (1986) also found a significant correlation between the techniques (*r* = .67), but it is not as high as that found by Vidulich and Tsang (*r* = .78). However, specific differences were found. A major difference was that the between-subject variability was consistently lower for the NASA bipolar ratings than for SWAT. It was suggested that even with the high level of concordance between subjects' rank orderings, the SWAT scale development still represents a group average. For the bipolar scales, however, the weighting procedure individualizes the workload score.

  The relative ease of use of NASA bipolar and SWAT were also compared. SWAT can be used in real-time data collection as it only requires choosing one of three levels for the three dimensions. The NASA bipolar scales require a break in performance to collect the ratings. However, the SWAT card sorting procedure takes at least 20 minutes and may take as long as one hour. It was suggested that the NASA workload parameter comparisons were easier to perform and require about 10 minutes to complete the 36 paired comparisons.

  Neither technique was able to detect resource competition effects in dual-task situations, in response execution processing demands, or in the dynamics of difficulty changes. It was not certain if this resulted from inherent limitations in subjective methods or in the limitations of these two techniques in particular.

- Vidulich and Tsang (1987) classify the AHP as a relative judgment method as opposed to the absolute judgments of workload that are obtained with NASA-TLX or a unidimensional workload scale. All three OWL scales were used in a single-axis compensatory tracking task with control order determining the level of task difficulty and visual or auditory presentation. All three OWL measures exhibited close agreement in discriminating the task variables, although the AHP showed the greatest validity (as measured by correspondence to performance) and reliability (as measured by test-retest correlations). However, topics of concern include how well relative judgments could be made across more varied tasks, as well as the possibility of operators forgetting details or creating and reporting their own hypotheses about task relationships. Vidulich and Tsang suggest that the AHP appears promising and further research should be pursued.

- Warr, Colle and Reid (1986) used both SWAT and Modified CH to obtain workload ratings in a laboratory setting for both a cognitive and a motor task, each with three levels of difficulty. A linear transformation of SWAT scores was performed to make them equivalent to the Modified CH scores (conventional rounding rules are assumed). No statistical evidence was found that the scales differed in sensitivity. However, both scales were found to discriminate between task difficulty levels. The authors point out that, although the scales were found to be equally sensitive to the task manipulations, the SWAT subscales might provide more diagnostic information in an applied setting.

- Wierwille, Casali, Connor and Rahimi (1985) describe a study in which 14 workload measures including two rating scales were evaluated using perceptual tasks in a moving-based flight simulator. The perceptual tasks involved seeing warning lights on the instrument control panel and responding via a pushbutton. Both the Modified CH and the WCI/TE rating scales were used to obtain workload ratings. Both rating scales showed a monotonic increase in ratings as the task difficulty increased across three levels. The scales differentiated between high and the other two levels of task difficulty. Little difference was found in the ability of the two scales to reflect changes in workload.

  Wierwille et al. (1985) also report a similar experiment using mediational tasks comprised of finding geometric and mathematical solutions to various navigation problems. Once found, the solutions were not implemented. The Modified CH and WCI/TE were used to obtain OWL ratings. The Modified CH showed a monotonic increase in workload ratings as difficulty increased while the WCI/TE showed no difference between low and medium difficulty. The Modified CH would therefore be recommended as the better rating scale for OWL assessment in mediational tasks.

- Wierwille and Connor (1983) evaluated 20 workload measures including two rating scales using a psychomotor task in a moving-based flight task simulator. Both the Cooper-Harper and the WCI/TE scales were used. The results indicate that both rating scales significantly discriminated between each of three levels of task difficulty. The normalized means of each difficulty level corresponded exactly in rank order and closely in magnitude. Both scales were found sensitive to and are recommended for workload measurement for psychomotor tasks.

- Wierwille, Skipper and Rieger (1984) conducted two studies to test whether the sensitivity of the Modified CH could be increased by changing from a 10-point to a 15-point scale or by changing the format to computer-based or tabular form. Increasing the categories from 10 to 15 did not consistently improve sensitivity. In general, they concluded that the original Modified CH was the most consistently sensitive measure of the five alternatives tested.

Two observations can be made regarding the comparative studies of subjective workload measures. The first observation is that, when the techniques are used for the same task, in general the results are very similar. In all studies using two or more different techniques (excluding Wierwille et al., 1984 and Hart & Staveland, 1987), the same rank order of difficulty was found for the task loadings. It appears that each of the techniques described, when carefully planned and implemented, can provide useful assessments of OWL.

The second observation is that more comparative work of this kind should be done. Following the traditional lead from psychometrics, it is believed that factor-analysis and other structural investigations would provide a stronger base for comparisons among techniques. Certainly, comparisons of the various techniques are required for systems applications of interest to the Army.

Issues Concerning Subjective Rating Techniques

Dissociation between Subjective and Performance Measures

Subjective workload measurement and operator performance are generally highly correlated during the early and middle stages of overload. Higher subjective ratings of workload are obtained in parallel with worse performance. However, this pattern is not always the one obtained in OWL assessments. For example, a dissociation between performance and subjective measures may occur where one task is performed better than another but is perceived as having higher workload (Yeh & Wickens, 1984). The idea of dissociation between subjective measures and performance is troubling because it indicates that opposite conclusions might be drawn, depending on whether subjective or performance measures are used for evaluation.

Although this continues to be an active research area, several conclusions of interest to practitioners have been drawn. In general, subjective experiences are more assessable via introspection and verbal reports when they are in working memory (Ericsson & Simon, 1980). Therefore, perceptual and cognitive elements (i.e., those elements associated with working or long-term memory) will be more salient in subjective reports than those elements that are associated with response execution such as control manipulation. Yeh and Wickens (1984) ran a series of experiments to investigate various hypotheses regarding dissociation. Based on the results, they conclude that the strongest dissociation occurs between single task difficulty and a dual task combination. When performing two tasks together, increased cognitive management is needed for processing and coordination. Therefore, subjective estimates will be higher than actual performance decrements in an easy dual task in comparison with a hard single task. Yeh and Wickens also found numbers of display elements increased the subjective workload experience although tracking performance was helped with the multielement predictor display.

Vidulich and Wickens (1986) make several observations on the implications of dissociation between subjective OWL measures and performance. One observation is that the usefulness of subjective measures may be reduced in detecting the individual workloads of single subtasks in a multitask environment. Therefore, the authors suggest, subjective OWL measures should be obtained on important subtasks in a single-task environment, if possible. Otherwise, perhaps the multitask environment differences should be weighed more heavily than those found in a single-task situation. Another source of dissociation is suggested to result from subjects' logical analysis of the situation. Vidulich and Wickens, for example, slowed down the presentation rate of stimuli, this disturbed the subjects response rhythm, and consequently degraded performance. However, the subjects' perceived the slower rate, and logically deduced that this should cause less workload, and based their ratings on that analysis. A third dissociation found by Vidulich and Wickens is that associated with increased motivation. Higher levels of motivation (induced by bonus pay) aided performance but led to perceptions of higher workload. The implications for operational settings include the importance of maintaining constant motivation levels for different subjects and tasks during system evaluation. The authors emphasize the importance of subjective measures in situations where new or interesting alternatives might influence performance and obscure actual differences in OWL.

Vidulich (1987) restates the observation that "subjective workload assessments are sensitive to manipulations that influence the perceptual/central processing demands and relatively insensitive to manipulations that influence response execution demands" (p. 8). Therefore, subjective measures are particularly useful in situations where operators are system monitors and the primary tasks are involved with perception and decision making.

Practitioners do need to be aware of possible causes of dissociation of subjective workload measures and performance. Several practical implications have been mentioned as well as suggestions for ways to handle them. The bottom line is that neither subjective nor performance measures should be used as the sole basis for assessment of OWL.

Delay of Ratings

The effect of delay in operators giving ratings has been briefly touched upon in previous discussions. The concern is that if operators are unable to give ratings in real-time, the passage of time and behavior that follows may affect subjective workload assessments which are made. Short delays of up to 15 minutes have not been found to have a significant effect on subjective ratings (Eggemeier et al., 1983; Hart et al., 1984), although some differences were exhibited when a difficult intervening variable was presented (Eggemeier et al., 1984).

Video recording the operators' activities can serve as an aid for collecting ratings in a post-test session. This method was used and determined to be a viable alternative to real-time ratings by Gidcumb (1985) when using SWAT. Although this method of video taping activities for post-test visual recreation has not yet been reported extensively in the literature, it appears to be a viable alternative when real-time ratings are not available due to safety or other practical constraints. Another alternative would be the use of the AHP technique to obtain relative comparisons of OWL during post-test sessions (Lidderdale, 1987; Vidulich & Tsang, 1987).

Relative vs. Absolute Measurements

There are two ways in which ratings can be used. First, several OWL ratings can be used in a relative sense to compare whether one task or activity has been perceived to have a higher workload than another task or activity. Second, absolute subjective OWL ratings are intended to indicate the level of workload without reference to any other task or activity. However, the question remains whether any subjective workload rating scale can be used to make absolute judgments. Subjective opinion is largely based on experience. As Lidderdale (1987) observes, "It is possible that all assessments of workload are made from a baseline of comparisons with other elements in the flight and, if this is the case, all rating methods may be relative" (p. 73). The absolute judgment of workload may be based on what has been experienced previously; the highest workload experienced may be the touchstone of what is considered high workload. If more difficult or intense tasks are performed, the touchstone for high workload may change. It is uncertain if individuals can possess an absolute scale for OWL that will remain stable over time.

Some OWL techniques are explicitly relative, such as magnitude estimation or the AHP. Other scales address the issue in a different way. The NASA scales, for example, ask operators to judge the relative importance of scale dimensions with respect to each individual task, thereby producing weightings for individuals by task. Individuals' card sorts of SWAT have been found to be relatively stable over time and the operational definitions of the levels of each dimension remain constant. It could be inferred then, that each individual would have an absolute workload scale. There is an anecdotal impression based upon such an inference that a SWAT rating of 60 or greater indicates a high workload condition. Certainly, a high rating on any scale should be pursued as a potential indication of OWL problems, although the relative nature of subjective ratings may lead to inappropriate interpretations and conclusions. The relative nature of subjective ratings also cautions against comparing systems evaluated in different studies (e.g., subsequent models of the same combat system).

Individual Differences

The issue of differences between individuals in the perception of workload is a continuing question of interest to researchers and practitioners. In the context of subjective measures, the issue is one of individual definitions of workload and what aspects of a particular task or activity are considered relevant to the assessment of workload. The NASA-TLX was designed to specifically address this issue by using individual weightings of the importance of each scale dimension to obtain a workload rating. This has been shown to reduce the between-subjects variation (Hart & Staveland, 1987; Vidulich & Tsang, 1986). The conjoint analysis used in SWAT seeks to account for individual differences through the scale development (i.e., card sort) procedure as well as the development of prototype scales (Reid, Eggemeier, & Nygren, 1982). The use of z-scores provides comparability between the widely varying scores produced via magnitude estimation. However, because of the intersubject variability, OWL evaluation must always use a sufficient sample of subjects, otherwise mean scores obtained in tests may not provide sensitivity.

OWL intersubject variability is also of concern because of its potential implication for selection of personnel to operate systems. Do differences between individual ratings reflect orderings of capabilities for handling systems? Unfortunately, there is a dirth of information concerning the interrelationships between individual differences in ratings of workload and information processing-related variables used by the Army such as the ASVAB. OWL individual differences are therefore an area for investigation because of the implications for the Army.

Questionnaires and Interviews

The second broad area of subjective methods are those that use questionnaires, interviews and other techniques to obtain estimates, judgments, evaluations, comparisons, attitudes, beliefs or opinions of people (Dyer, Matthews, Wright, & Yudowitch, 1976). Such methods are frequently used and are seen as useful (Meister, 1986). The major reason for the widespread use of such methods is that they are perceived as easy and quick to administer, particularly in field test environments, and inexpensive to develop and produce.

Questionnaires

Questionnaires are forms in which written questions are asked in a fixed order and format and to which respondents write their answers. The questions may be open-ended, allowing respondents to write in their own words and make any answer, or close-ended, where the choice of answers has been previously established, such as multiple choice or true and false. Meister (1985) states that the results of studies (Ellenbogen & Danley, 1962; England, 1948; Kohen, de Mille, & Myers, 1972; Prien, Otis, Campbell, & Saleh, 1964; Scates & Yoeman, 1950) suggest that open-ended questions may provide unique information, but close-ended questions are more reliable. A number of sources are available for guidance in the development of questionnaires, including the advantages and disadvantages of various types of questions, sequencing and wording of questions, etc. (Dyer et al., 1976; Meister, 1985; U.S. Army Test and Evaluation Command, 1975).

The development of useful questionnaires requires not only the choice of question types and proper wording, but also the content of the questions -- What do they ask? They need to be designed to obtain the desired information. Pretesting of questions to ensure their appropriateness to the desired end as well as planning of the data analysis are important to a questionnaire's value.

The advantages of questionnaires are that they are less expensive and can be completed faster than interviews or ratings. Questionnaires often can be handed out and collected without attaching names to the answers; hence, they can be more anonymous than interviews. As a result of their anonymity, questionnaires may garner more self-revealing and unfavorable reports than interviews which rely on one-on-one communication.

There are problems associated with the use of questionnaires in test and evaluation environments. System experts may devise the questions and not have expertise in question development. Yet, as a result of the ease of putting together a question, there is a tendency for questionnaire use to proliferate. An example given by Taylor and Charlton (1986) describes widespread respondent burnout from having answered too many questionnaires that were too long and not focused on the operator's activity. The end result was a vast collection of meaningless data. The frequent use of not well-thought-out questionnaires can result in data that are large in quantity but limited in usefulness.

A recent move has been in the direction of creating computerized systems to create well-constructed, focused questionnaires for specific purposes. Enderwick (1987) describes a system where a catalog of well-crafted questions on human factors test and evaluation topics is available to operational test directors (OTDs) who may choose any number of applicable questions. The questions are printed out (with the name of the equipment substituted for the word "equipment" so that the questions appear designed for that test) and can be re-ordered by OTDs. The final package is printed out with a cover page, an instruction page and the questionnaire. The computerized questionnaire system is designed for use by people who did not necessarily have any human factors training. Test directors could design questionnaires to meet their needs and new questions could be added.

Taylor and Charlton (1986) have developed an automated adaptive questionnaire which used a branching concept to determine what questions will be asked contingent on the answers to previous questions. The respondent answers general questions on a seven-point scale and if the answer meets some predetermined criterion (e.g., -2, with -3 being the most negative score), more detailed questions will be asked. The contingency branching method is most suitable for computer implementation. Computer implementation also allows on-site data analysis of answers. Note that questionnaire procedures may incorporate a scaling method thus blurring the distinction between rating scale and questionnaire.

Questionnaires are commonly used in test and evaluation environments (Enderwick, 1987; Meister, 1986). Anecdotal evidence indicates they are commonly used for workload assessment although the specific questionnaires are rarely found in the research literature. It certainly appears that the development of workload questionnaires aided by computers could be helpful to Army analysts. It seems that sufficient information is currently available to create a universal set of general workload assessment questions that can be tailored for specific application. However, further development of this concept is needed before such a tool is available for use.

Interviews

The interview is an interpersonal interaction in which the interviewer seeks information or opinions from the respondent. It permits more flexibility than the traditional questionnaire. It allows the interviewer to follow-up on the answers given and thereby gain insight into areas that may not have been addressed in a written questionnaire. Disadvantages of the interview method are that it is very costly in time and, because of the personal communication, respondents may be less likely to report anything negative and may also be influenced (consciously or subconsciously) by the interviewer. Interviews can take place one-on-one or with groups as in the case of crew operations. Key questions can be determined ahead of time and pretested for understanding, likely responses and the information they contain (Meister, 1985). Both Dyer et al. (1976) and the U.S. Army Test and Evaluation Command (1975) provide information about interview considerations, procedures and analysis. Interviews are useful to obtain unique information and opinions about workload. Meister (1985) suggests that test participants should always be interviewed to learn how the participant viewed the test situation. If the view was different from that intended by the test director, then the data may have been affected.

Protocol Analysis

Protocol analysis requires operators to verbalize their thought processes or performance. This method has been used extensively in computer interface research as a way to find out how an operator solves problems or discovers the appropriate commands to use. Protocol analysis relies on the ability of the subject to determine introspectively thought processes and then verbalize them, either during task performance or afterwards. Verbal protocol is listed as an available subjective workload method (Hart, 1986a). Brown (1982) writes that such verbal reports can be very informative, but during high workload, operators may not have the time to provide complete information. (In a sense, the verbal report is a secondary task.) Verbal protocols and analysis may provide useful information, particularly in computer interfaces where such techniques have previously been used.

Summary

Questionnaires and interviews provide an important adjunct to workload estimation. Proper questioning can provide insight into the causes of problems associated with workload. Furthermore, questionnaires and interviews provide an opportunity for subjects to give their detailed impressions of system operation and how it might be improved. Rating scales are usually too highly structured to provide detailed, subtle impressions. Questionnaires and interviews require careful construction and should be used to obtain more detailed information in all workload assessments. Possible enhancements to these subjective measures are an automated questionnaire design tool and the use of protocol analysis.

Summary and Conclusions

The need for subjective techniques for workload assessment in applied settings has been identified and substantial efforts have been directed toward obtaining a solution as evidenced by the amount of research performed and reported. Several recommendations can be made based on the review and analysis of the subjective techniques and the issues involved in their use:

- Subjective measures can provide valuable information concerning the operators' perception of their workload experience in specific tasks or activities. Subjective techniques have been demonstrated to be sensitive and should *always* be included in an evaluation wherever possible.

- The questions of interest to the system designer or evaluator should be defined before choosing a technique by which to obtain answers. Overall workload ratings, such as the Modified CH, will provide a global assessment and can identify potential problems or workload chokepoints. More specific information, like that available through multidimensional scales or questionnaires and interviews, will be necessary to potentially diagnose specific sources of workload and identify solutions.

- The value of qualitative information, like that obtained from questionnaires or interviews, should not be underestimated.

- All subjective measures, including questionnaires and interviews, must be carefully planned and implemented to obtain valid and useful data.

- The OWL evaluator should be aware of the measurement scale characteristics (ordinal vs. interval; uni- vs. multi- dimensional) and to what extent these characteristics will influence the interpretation of results and conclusions that can be appropriately drawn.

- Multidimensional scales, like NASA-TLX and SWAT, offer the opportunity for using the subscale ratings in diagnosing OWL with respect to specific system design characteristics.

- Available evidence indicates that Modified Cooper-Harper, NASA-TLX and its predecessor, and SWAT are sensitive to differences in workload. Substantial research supports their use in OWL assessments. Less information is currently available on the Bedford scale and AHP. The original Cooper-Harper scale has been found to be particularly sensitive to psychomotor tasks in aircraft environments. It is not known if it is equally sensitive to psychomotor tasks in other system control activity (e.g., tank operation).

- Psychometric scaling techniques have been shown to be sensitive to differences in task manipulations. These are viable alternatives although a certain degree of knowledge concerning these techniques is required in order to meet necessary design, implementation, and mathematical requirements.

- The use of observers as well as operators to make OWL ratings is an alternative in workload assessment, although trade-offs in information quality exist.

Subjective OWL assessments can provide useful and valid information for the Army if there is: (a) careful definition of questions to be answered; (b) careful selection of technique; and (c) careful, consistent implementation of technique in a laboratory, simulator, or field environment.

CHAPTER 6. SECONDARY TASK TECHNIQUES

An important reason for measuring operator workload derives from the objective of designing human-machine interfaces that will optimize system performance. To do so requires knowledge of the work capacities and limitations of the human operator.

Secondary task techniques have been employed as a tool to assess the work capacities and limitations of the human operator with respect to primary task performance. Typically, the secondary task paradigm is used in applied settings to assess the workload associated with a primary task such as piloting an aircraft. To derive the workload associated with the primary task, the operator is required to perform an additional or secondary task simultaneously with the primary task. The relative workload associated with the primary task is reflected in the levels of performance on the secondary task. That is, because primary task performance requires the utilization of the resources and capabilities of an operator, secondary task performance will reflect the remaining resources and capabilities or relative spare capacity of an operator. For example, if the operator is fully loaded by the primary task, performance on the secondary task may be unacceptable. By contrast, if the operator is only partially loaded by the primary task, performance on the secondary task should be acceptable. (See Chapter 2 for a description of the relation between human performance and operator workload.)

A critical aspect of the secondary task paradigm is the determination of acceptable and unacceptable performance on the secondary task. This determination may be accomplished by establishing the performance level on the secondary task without the primary task, and then comparing this baseline performance level to secondary task performance with the primary task. The determination may also be accomplished by varying the difficulty of the secondary task while the operator maintains the primary task performance. Then the comparison is on secondary task performance across the levels of difficulty. Through these various manipulations, the secondary task paradigm offers the practitioner a means to assess the relative workload associated with a primary task which may not be apparent from primary task measures alone.

Secondary Task Paradigm: A Solution or a Problem?

The secondary task paradigm encompasses several techniques that have been employed to assess the spare capacity and resources available for additional work when performing a primary task. There have been many reviews on this topic over the past 25 years. For example, Knowles (1963), O'Donnell and Eggemeier (1986), Ogden, Levine, and Eisner (1979), Rolfe (1971), and Williges and Wienville (1979) have all provided reviews of the techniques as well as the methodological issues associated with their usage. Because secondary task techniques have received a considerable amount of attention, it would seem appropriate to assume that guidance in the use of such techniques would be straightforward and readily available. This is not the case.

To illustrate, Gopher and Donchin (1986) and O'Donnell and Eggemeier (1986) disagree with each other in the same volume of the *Handbook of Perception and Human Performance* concerning the methodological issues that should be addressed in implementing a secondary task technique. Specifically, O'Donnell and Eggemeier support claims that the secondary task must not interfere with the performance of the primary task. By contrast, Gopher and Donchin take exception to such a position and support the position that it is legitimate for secondary tasks to interfere with performance of the primary task.

For the practitioner concerned with operator workload, such mixed messages regarding the implementation of secondary task techniques are troublesome. In fairness to the authors just cited, their apparent disagreement illustrates the differences in opinion found in the literature in how best to select, implement, and interpret secondary task measures. The reasons behind such disagreements are based on: (a) theoretical grounds, (b) the findings from the plethora of secondary task studies, and (c) practical considerations. These are briefly discussed below as background for our approach for use of secondary tasks.

Theoretical

Measure Spare Capacity. One theoretical position has espoused the secondary task paradigm as a tool to provide an uncontaminated measure of the spare capacity or resources not expended with a primary task (Kahneman, 1973). This is the view of O'Donnell and Eggemeier (1986). Such a theoretical position requires the primary task performance to be stable when the secondary task is concurrently performed with the primary task. Then and only then can changes in secondary task performance be interpreted as a reflection of spare capacity leftover from primary task demands. (See Kantowitz [1985] for a critique of the spare capacity concept and the problems associated with such a concept as it relates to measures of performance.)

Load the Operator. A different theoretical perspective is to view the objective of the dual task paradigm to be the measurement of the operator's ability to perform adequately two tasks concurrently (Schneider & Fisk, 1982). This is the view of Gopher and Donchin (1986). This theoretical position maintains/argues that changes in primary task performance when a secondary task is performed concurrently reflects an inefficiency in human performance in the dual-task situation as opposed to a methodological flaw.

Wickens' Resource Model. Another important theoretical formulation is Wickens' Resource Model (Wickens, 1980). The resource model has been offered as a guide for secondary task selection with respect to the nature of such tasks (O'Donnell & Eggemeier, 1986). The model depicts the overall human information processing system as composed of multiple but separate processing structures/resources, each of which can have capacity limitations and be a potential bottleneck in the human processing system. These separate processing structures are defined along the following three dichotomous dimensions:

- stages of information-processing (perceptual/central-processing operation vs. response selection and execution),
- modalities of perception (auditory vs. visual), and
- codes of information processing and response (spatial-manual vs. verbal-vocal).

Each processing structure has its own limited supply of resources which are not interchangeable with other processing structures. It is suggested that the secondary task be selected so that it has the same processing structures utilized by the primary task. In this manner, the secondary task is more sensitive in identifying the level or amount of spare capacity (O'Donnell & Eggemeier, 1986). In support of this position, Shingledecker, Acton and Crabtree (1983) conducted a study that demonstrated that the sensitivity of the secondary task performance varied as a function of the primary task resource demands according to Wickens' model. Such results are promising, but it may not be readily apparent which processing structures are dominant with performance on a complex system such as a helicopter.

Summary. Therefore, depending upon one's theoretical position and primary interest in using the secondary task paradigm, the selection of a particular secondary task technique will vary. As a result, secondary task selection in applied settings may still be difficult.

Results of Studies: What to Believe?

Another contributing factor to the apparent confusion concerning the appropriateness of various secondary task techniques stems from the difficulty of simply interpreting reported findings. For example, Ogden et al. (1970) provided a table containing 144 secondary task studies and listed the major findings from these studies. Perusal of the table reveals that for most secondary tasks one study can be cited to show improvement, another degradation, and a third no change in secondary task performance. It is consequently not readily apparent which secondary tasks are most appropriate to assess OWL. Appendix A contains a detailed review of the secondary task literature which illustrates this complexity.

Practical Considerations: Research vs. Applied Context

The vast majority of the work done with secondary tasks has been conducted in controlled laboratory situations. As a result, secondary tasks that have been found sensitive and capable of measuring spare capacity in laboratory situations may not be applicable in applied settings. For safety reasons, for example, flying a helicopter precludes the use of any secondary tasks that would possibly interfere with the pilot's ability to maintain control over the aircraft. Another practical consideration is that the elaborate experimental procedures usually required to implement a secondary task paradigm may be excessive for many system development efforts in terms of manpower and time constraints.

The remaining sections of this chapter consider secondary task techniques as used in applied workload assessment settings. It differs from other chapters in that we refrain from reviewing all the literature at this point for two reasons. First, there is a tremendous volume of literature. Second, most of the literature is theoretical and academic in nature. Although much of the discussion in the literature is of considerable importance in understanding how cognitive components of workload impact human performance per se, it may be of secondary importance to the individual evaluating a system. As an alternative, we have opted to put the more general review of the literature in an appendix (Appendix A) so that it is available for the interested reader. We will now discuss our approach and then we will present examples concerning design issues and suggest applications of secondary tasks.

Our Approach

The problems facing the practitioner interested in workload assessment are: (a) knowing the circumstances in which secondary task techniques are appropriate, and (b) which ones to use. Our approach is a systematic attempt to provide such answers in identifying appropriate secondary task techniques within the context of a system development effort. The approach is directed from a very pragmatic philosophy. That is, secondary tasks offer utility for a system development effort when such tasks are used to load the operator and drive him to the performance envelope boundary. The purpose is to determine how much more can the operator do. This chapter describes the most practical secondary task techniques that allow such measurements.

To evaluate the appropriateness and utility of secondary tasks in applied settings, it was deemed necessary to examine the specific design issues or concerns that would call for using secondary task techniques. It is important to recognize that the basic secondary task paradigm encompasses several different techniques which manipulate or vary the parameters of secondary tasks in order to identify potential OWL problems with a primary task. We judge these various techniques for their appropriateness in answering specific design questions by reviewing the literature in support of such techniques. The utility of secondary task techniques was also examined with respect to meeting the Army need for workload techniques that are relatively easy to implement, can be used to identify OWL problems within complex systems, and provide relatively straightforward application for data collection and analysis.

Secondary Task Techniques in Applied Settings

Previous reviews noted that secondary task techniques are most applicable to early design stages of systems in controlled laboratory settings (e.g., Schiflett, 1976; Wierwille & Williges, 1979). Several factors have been suggested for their lack of use or applicability during the later phases of the system development process (Ogden et al., 1979; Shingledecker et al., 1980). For example, Ogden et al. (1979) noted that secondary task techniques may not receive uniform operator acceptance. As a result, operators possibly will run the gamut from neglecting the secondary task altogether to assigning it such a high priority that it artificially contaminates and changes the test situation. In either case, the results from such test situations will not accurately assess the amount of resources committed to the primary task or the amount of spare capacity remaining.

More recently, researchers have suggested the applicability of some secondary task techniques for use in simulations as well as the later phases of system development (e.g., Bortolussi, Kantowitz & Hart, 1986; Shingledecker, 1987). These techniques are designed to alleviate problems such as:

- instrumentation limitations which preclude the use of secondary tasks into system prototypes or high fidelity simulators,
- potential task intrusion caused by the use of secondary tasks, and
- poor operator acceptance of secondary tasks (Shingledecker, 1987).

Such techniques offer great promise for the Army since they seem to overcome the potential objections concerning operator acceptance and artificial intrusiveness on primary task or system performance. In addition, these techniques are relatively easy to implement. Four specific design and development examples in which these secondary task techniques offer the greatest utility are described in subsequent sub-sections. For each of the examples, a brief description will be given and subsequently discussed with regard to appropriate secondary task techniques. These discussions are intended only to provide sufficient detail for understanding secondary task techniques. Following discussion of the examples, consideration of other potential secondary tasks will be given. Our integrated approach to a workload assessment battery containing several different types of techniques is discussed in Chapter 8.

System Design and Development Example 1

Description. Successful operation of a system requires that the operator routinely perform several tasks in order to carry out a mission (e.g., tracking targets, radio communications, weapon delivery, etc.). You are interested in knowing whether an operator can adequately perform these tasks. Specifically, are there limits in the operator's capability to perform these tasks, such that if the limits are exceeded the operator's performance deteriorates? (This is an example of overloading the operator, although experiments do not always show reduced performance on the primary task.)

Considerations for Secondary Task Techniques. The *embedded secondary task* technique developed by Shingledecker and colleagues (Shingledecker, Crabtree, Simons, Courtright & O'Donnell, 1980; Shingledecker & Crabtree, 1982) offers a means for such an assessment. The concept of the embedded secondary task is based on overcoming the problems of implementation, intrusiveness, and operator acceptance mentioned earlier. The embedded secondary task technique alleviates these concerns by utilizing an existing sub-task of the system, such as radio communications as the secondary task, that is fully integrated with existing system hardware and software and with the operator's conception of the mission environment. To illustrate, Shingledecker and Crabtree (1982) reported a study in which they used the radio communications task in an aircraft environment as the embedded secondary task. They scaled the various task loading properties of several radio communication messages. The task load is the work pilots are required to perform in response to such radio messages such as request for radio frequency change or request for traffic information. Based on their scaling of radio message task load, they were able to infer that increased communication load produced decrements in the primary task performance of operator tracking in a flight simulator. Similarly, radio messages that were more demanding also elicited signs of overload in secondary task performance; the operator took longer to perform required actions in response to such messages when compared to control conditions, i.e., the radio task by itself. Such findings encourage the use of embedded secondary tasks in assessing the limits of operators' workload capabilities.

Similar findings have been reported in several simulation studies (e.g., Wierwille, Casali, Connor, & Rahimi, 1985). The primary task in the studies reported by Wierwille et al. (1985) required pilots to maintain a steady course under simulated conditions (e.g., mild random crosswind). Within each study, a task that can be described as an embedded secondary task was manipulated to increase the demands on the operator by the tasks. For example in one study, they varied the number of warning and emergency lights that pilots were required to detect (monitoring task). In another study, they varied the complexity of wind-triangular course problems (navigation tasks) to be solved during the simulated flight. In a third study, they varied the number of occurrences of the pilot's call sign (radio communications task) to which pilots responded. Wierwille et al. (1985) do not classify these manipulations of task parameters as embedded secondary tasks, although their use of such sub-tasks fits the embedded secondary task paradigm.

The results from the Wierwille et al. studies were quite revealing. In all cases, the manipulation of the embedded secondary task demands resulted in reduced pilot performance on secondary tasks, while the performance measures for the primary task remained relatively stable. Such findings are indicative of pilots' capacity to handle workload demands (i.e., spare or reserve capacity as assessed by embedded secondary tasks). These results are further substantiated by the fact that more traditional secondary task techniques, time estimation, were also used in the study and exhibited similar results.

Finally, Chiles and Alluisi (1979) with a multiple-task performance battery (MTPB) have used similar logic as employed in embedded secondary task paradigm. In particular, they assumed that the monitoring tasks in their task battery were acting as secondary tasks. Based on this assumption, they used the monitoring task results as indices of workload imposed by different combinations of the other, time-shared, active primary tasks to develop a workload metric. Taken together, the body of these results lead to the conclusion that the workload associated with time-shared multitask systems can be assessed by use of the embedded secondary task technique.

System Design and Development Example 2

Description. You have two alternative designs of a system or sub-system which have been shown by previous testing to be essentially the same (no differences) with respect to primary task measures. In this situation, you are faced with what appears to be two comparable designs. Which system design do you choose? (This example is one in which the practitioner might like to determine where the operator is in the workload performance envelope. However, this determination is not absolutely essential to answer the question.)

Considerations for Secondary Task Techniques. Besides the potential cost factor differences between the two designs, there may also exist operator workload differences that are not being reflected by primary task measures. The secondary task paradigm can be used to determine if either of two designs is less demanding on the operator. This is important because the less demanding design will leave more spare or reserve capacity so that the operator can perform the mission tasks under more demanding conditions (e.g., combat) than those investigated.

Embedded Secondary Task. The embedded secondary task technique is applicable to this design example if the two alternative designs have subtasks that can be used as the secondary task. In fact, Shingledecker (1987) describes a situation similar to the design example described above in which the embedded secondary task technique is offered as the vehicle to identify the most appropriate design alternative.

If the embedded secondary task technique cannot be applied, there are several traditional secondary tasks which may be useful. These secondary tasks have been used to determine the spare capacity of operators when engaged in complex system operations such as flying an aircraft.

Time Estimation Secondary Task. Typically with this task, operators are required to produce time intervals of 10 second durations without using counting, tapping or any sort of direct timing procedures (see Hart, 1978, for the merits of such a time estimation procedure). The premise behind this procedure is that the busier an operator is, the less attention is available to judge time accurately and as a result the subjective impression of time becomes less accurate with respect to objective time. That is, operators will produce longer and more variable estimates of 10 second time intervals because they lose track of time. To illustrate the use of this technique, Bortolussi, Kantowitz and Hart (1986) conducted a study with pilots in a Singer-Link GAT-1 flight trainer. Two full-mission instrument-flight-rule scenarios (high and low workload scenarios) were utilized. In addition, each scenario was designed to contain flight segments that varied in difficulty. The results were such that the time production secondary task discriminated between low and high workload scenarios (i.e., longer time intervals for the high workload scenarios). Furthermore, it discriminated among individual flight segments in the high workload scenario but did not in the low workload scenario. The variability of time productions was also greater for the high workload scenarios than for the low workload scenario. Similar results have been reported by other researchers. For example in a series of studies, Wierwille et al. (1985) found the variability of time productions (i.e., standard deviation) discriminated between various workload conditions that were manipulated within a flight simulator. The workload conditions involved task loadings or workload levels on either psychomotor, perceptual, mediational, or communication task components of the flight simulator. The merits of using the time estimation production technique for assessing the relative workloads of two comparable design alternatives are several. It requires little instrumentation or training and can be included as a normal part of an operator's duties without interfering with such duties (Hart, 1986).

Choice Reaction Time Secondary Task. Another secondary task that is relatively easy to use is choice reaction time. This technique involves operators responding to several visually presented stimuli (e.g., a light emitting diode with arrows pointing in different directions), with each stimulus requiring a different response such as different buttons to press. To illustrate, the Bortolussi et al. study (1986) cited above also included 2 and 4-choice reaction time tasks that pilots performed during the flight scenarios. Mean reaction time scores for both 2 and 4-choice reaction time tasks discriminated between low and high workload scenarios. The choice reaction time tasks also discriminated the different workload levels among different flight segments. These results have been replicated in another study by Bortolussi et al. (1987). The merits of using choice reaction time as a secondary task lies in its simplicity, ease of implementation, and ease of interpretation of results. Moreover, its sensitivity follows the theoretical basis for its use as a secondary task; that is, it reflects central information-processing demands as well as response selection demands.

System Design and Development Example 3

Description. You have a system that is under the Product Improvement Program (PIP) for enhancements or modifications. You are interested in whether the operator can handle the new capabilities and/or new functionality that is planned.

Considerations for Secondary Task Techniques. If the specific enhancement is definable as a new subtask, it can be examined easily within the framework of the embedded secondary task technique. *The new task can act as a secondary task.* The demands (e.g., the timing and number of radio messages received with a communications task) associated with the new task can be varied to examine its effects on the operator's performance with the existing system. By employing the embedded secondary task technique, it is possible to elucidate the conditions under which the new task may, in fact, hinder operator performance. Another variation of the embedded secondary task technique would involve setting the new task aside and the manipulation of an existing subtask of the system as the secondary task in order to determine the limits of operator performance. By so doing, it is possible to estimate the spare capacity an operator would have for a new subtask.

If these variations of the embedded secondary task technique cannot be applied, there are several other secondary tasks which may be useful. Scenarios for system usage can be developed within which time intervals can be identified for the operator involvement with the new task. A secondary task can be substituted for the proposed task to examine the spare capacity that would be available to perform the new task within the context of the system's other requirements (tasks) placed on the operator. Choice reaction time and time estimation are two secondary tasks that may be applicable for these circumstances. Bortolussi, Hart, and Shively (1987) provide evidence for the use of secondary tasks in a synchronized manner with specific scenario events in order to identify changing workload levels within the context of a flight simulator. They synchronized the presentation of a choice reaction time task and time production interval task to specific events during high and low workload flight scenarios. By so doing, they were able to discriminate with both secondary tasks between high and low workload scenarios. They suggested that these results could be further examined by a detailed time-line analysis to localize the specific events that produced the apparent differences between flight workload scenarios. This is similar to the proposed use of secondary tasks being offered in this section.

System Design and Development Example 4

Description. You have a system under test and evaluation. You are not only interested in knowing whether the system can be handled by operators within the context of a mission scenario but also where the potentially high operator demand, areas lead to operator workload problems. Clearly, loading the operator will show performance deficiencies that identify the high workload areas.

Considerations for Secondary Task Techniques. It is quite probable that primary task measures will answer the direct question concerning the operator's capabilities to handle the system within a set of conditions tested. With respect to identifying the areas that are relatively high in workload demands, the embedded secondary task paradigm can be utilized. The supposition is that the operator can be driven to performance limits by various task loadings on the designated secondary task. By so doing, breakdowns in human performance can be identified that may otherwise not be shown with primary task measures under normal circumstances. Additionally, if there is a possibility to break the mission into segments, examination of the performance within segments will help to identify the problem areas.

Another possible method is to synchronize secondary task presentations to specific primary task sequences that may be suspected of high workload but may not be reflected by primary task measures. You are, in essence, attempting to identify momentary high workload areas that may under stressful circumstances contribute to poor operator performance. The Bortolussi et al. (1987) article described above is an example of using secondary tasks in this manner. Based on this study, choice reaction time tasks and time interval production tasks may be appropriate for such a type of operator workload assessment.

Other Secondary Tasks

With respect to secondary task techniques not specifically described in this chapter, there are several that have been shown to be sensitive to operator workload levels. For example, the Michon Interval Production Task (IPT) requires the subject to generate a series of regular time intervals by executing a motor response such as a finger tap every two seconds (Michon, 1964). The IPT has been shown by Shingledecker et al. (1983) to be sensitive to psychomotor task loadings for primary tasks but not other types of task loadings such as memory and perceptual sustained attention. Accordingly, the Michon paradigm seems appropriate for assessing psychomotor workload. What limits its applicability is the fact that the operator must perform the IPT with one hand devoted continuously to the task and as a result may limit its use with complex systems that require operators to have free use of both hands.

The Sternberg Memory Task (Sternberg, 1966) has also been shown to be sensitive to operator workload levels (e.g., Spicuzza, Pincus, & O'Donnell, 1974). The task involves memorizing a set of items, usually digits. Later in testing, a single probe digit is presented. The operator's task is to indicate whether the probe was in the memorized (positive) set. Both response time and accuracy are measured. Memory load may be varied by including different numbers of items in the memory set. Research shows that reaction time to the probe increases linearly with the number of items in the memorized set. In this way, a slope can be determined which reflects the rate of memory search and the degree of cognitive loading. Wickens, Hyman, Dellinger, Taylor, and Meador (1986) reviewed seven studies that employed the

Sternberg task in flight simulators or aircraft environments. The power of the Sternberg task lies in its potential diagnostic value in distinguishing between cognitive processing task loading and response task loading for a primary task. This requires analysis of the Sternberg task data for changes in the slope and intercept of such data in order to infer task loadings on either cognitive processing or response selection as a result of primary task demands. Based on their review, Wickens et al. (1986) have questioned the utility of such an analysis in a typical operator workload assessment situation since the studies reviewed reported a high degree of instability in the slope and intercept data. As a result, Wickens et al. (1986) have recommended the use of one level of the Sternberg task as a general memory secondary task to infer operator workload levels. Wickens et al. (1986) have also noted that the Sternberg task may be insensitive to high workload levels because pilots may shun the task under high workload.

Conclusions

Of all secondary task techniques, the embedded task offers the most practical utility for the Army. By utilizing this technique, one may overcome many of the problems identified in using these largely laboratory oriented secondary tasks in applied system evaluation environments. The principle advantage of the embedded secondary task technique is that the data collected are generally applicable with respect to design and system evaluations.

The other secondary tasks offered in the examples are possible alternatives that may be applied when the embedded secondary task technique is not feasible. However, these other techniques are offered with two cautionary notes. First, as shown in Appendix A, all secondary tasks can sometimes intrude on primary task performance. This possibility cannot be ruled out for the secondary tasks described in this review. However, the secondary tasks recommended here are ones that have been shown to minimize this potential confounding in most situations. A second consideration for these alternative secondary task techniques is their sensitivity to reflect primary task demands (i.e., workload). That is, these techniques may not always be applicable for a particular situation. They have been shown to be sensitive to workload levels in complex aircraft systems, but have not been fully exercised with other types of complex systems of interest to the Army.

The recommendations offered should not be interpreted to mean that other operator workload techniques are inappropriate in the circumstances described (e.g., subjective techniques). Indeed, we recommend that secondary task techniques be utilized as part of a battery of both subjective scales and other empirical methods. In this way, the information obtained from several diverse techniques can compensate for limitations in each individual method. Chapter 8 amplifies upon the breadth of these other techniques that are also appropriate for the situations described above.

CHAPTER 7. PHYSIOLOGICAL TECHNIQUES

Physiological techniques assess the operator's workload in a way different from primary and subjective techniques. Primary techniques sample directly observable responses of the operator, the operator's behavioral output resulting from some task. Workload (the relative capacity to respond) is inferred from the number, latency, or pattern of the responses. Subjective workload techniques assess the judgments of the operator about workload. These judgments are related to and directed toward such factors as frustration, difficulty, time pressures, etc. Workload is inferred from the judgments. By contrast, physiological techniques assess activities which are normally not directly observable and represent indices of the underlying processes involved in responses.

The OWL physiological literature has a well developed empirical, statistical and mathematical foundation. Further, many quite different techniques have been used. However, many authors often assume, apparently, that the reader understands the underlying physiology and the authors do not present the thinking and physiological rationale behind the application. When this happens, the various techniques used may ~~appear to be~~ an apparent 'grab bag' of techniques. In fact, the various techniques sample a range of quite different physiological systems and mechanisms. By and large, all of the techniques are based on sound physiological evidence, however, some techniques can be highly specific to a single physiological subsystem. Accordingly, it will be helpful to the reader to discuss the physiological basis and delineate not only the rationale, but also the physiological systems and mechanisms being measured.

In order to understand the application and the results obtained from such techniques it is necessary to caste each technique into an appropriate physiological context. First, we will discuss some measurement issues including the theoretical basis for each measure and some data analysis issues. Next, we will discuss briefly some basic physiology to provide a framework for the techniques discussed and especially what they measure. Then, we will review some of the literature on these techniques as the techniques apply to workload. Physiological measures of workload have been recently reviewed (e.g., O'Donnell & Eggemeier, 1986) and our intent is not to repeat the information already available but to build on it. Although results will be included, considerable emphasis will be on understanding of the technique and its application and usefulness in the workload context.

Preceding Page Blank

Reproduced From
Best Available Copy

Measurement Criteria

There is a major difference in the focus of physiological research and workload research and evaluation. The physiologist is interested in mechanisms and the functioning of physiological subsystems. Accordingly, the study of a physiological subsystem will involve discovering its full range of operation, including the impact of extreme conditions not normally encountered in every day life. By contrast, the workload researcher is interested in relating measurements of the subsystem under more normal circumstances to measures of human behavior and to workload. As an example, consider pupil diameter. The full range of pupil function is from about 1 mm in bright light to about 8 mm in total darkness and this range is what a physiologist would be interested in. Under normal circumstances, those which an operator is likely to experience, pupil diameter changes show a range less than 1 mm (Beatty, 1982). Clearly, the range of operation is much more constrained in the workload context.

When one uses a physiological technique to assess OWL there are several types of questions to answer.

- Is this physiological technique one that could reflect workload changes? This is an issue of **appropriateness**. Is the technique appropriate to the questions being asked? A large portion of the inconsistency about a technique in the workload literature may be related to this very point.
- Are the variations of the physiological technique in the normal operating environment sufficient to produce measurable workload variations? This is the issue of **sensitivity**. Are the techniques sensitive to the variations in OWL the operator will experience?
- Do these techniques reflect the kinds of changes in the human operator one is interested in? This is **diagnosticity** which is an extension of sensitivity. Are the techniques employed sufficiently specific to localize the difficulty and identify the underlying mechanism?

In one form or another, the techniques we have classified under physiological are those which are indirect indicators of operator workload as compared with primary task and subjective methods. They are presumed to be reflective of the amount and difficulty of the work the operator is doing. This is thought to be true because the bodily states vary as a function of what one is doing: waking, sleeping, running, sitting, etc. Similarly, changes in bodily state and especially brain states can be measured and related to these activities. It is the hope of the investigator that the bodily functions will show similar, measurable changes, albeit smaller for workload changes than what the physiology studies show.

Appropriateness

Early arousal theory assumed that all motivation and emotion involve the same basic continuum of physiological activation and that this continuum is reflected in all of the psychophysiological techniques (e.g., Hebb, 1955; Malmo, 1959). Thus, techniques of electroencephalogram (EEG), electromyography (EMG), skin conductance, etc. should be interchangeable according to arousal theory. It is now known that such a simplified view is wrong, but to an extent, the workload literature continues to reflect this simplified view. Increasingly, researchers are discovering workload to be multifaceted and thus a particular technique may reveal workload effects in one case but not another. Stated another way, the use of an inappropriate technique may be misleading; it may be a good technique in some instances, but it was the wrong tool for the question at hand.

As the task for the operator changes, the most appropriate workload technique for assessing OWL may also change. Many physiological techniques have been applied to the study of workload. In some cases, the authors have claimed the technique to have relevance for workload but the relevance is also contingent on the definition of workload. Because the nature of operator tasks has changed rapidly, some of the older techniques are more related to fatigue than to workload as defined in Chapter 2. These have been discussed, briefly. The appropriateness of the technique in the current workload context will be made apparent in the discussion of individual techniques.

Data Analysis Affects Sensitivity and Diagnosticity

There are several important implications of data analysis with regard to sensitivity and diagnosticity. Since mental workload is dynamically changing over time, the investigator should plan the study and the analysis to assess the timeline of operator activity. For instance, if one averages over time, one should be sure that the data have been examined first for consistent trends which may occur with time; these trends may be linear or nonlinear, depending on the circumstance. More will be said about trends in the section on heart rate. While this is not a caveat against averaging, it is a plea for knowledgeable and careful averaging. Unfortunately, there are cases in the literature that report a failure to find an effect but appear to have masked the effect of workload through averaging.

A fairly simple, preliminary analysis is to pick some (arbitrary) short time period as a window within which data are arranged. A computer program can average scores within each window to produce running averages. This permits the investigator to look for both short and long term trends in the data. The data can be reanalyzed using several variations including (a) changing the size of the temporal windows or (b) using temporally overlapping or non-overlapping windows. A considerable amount of information may be gained from analyses of this type; specifically, sensitivity will be increased.

Most applied workload studies involve human performance over a period of time. Time, then, is an especially important factor in the analysis of data obtained with the techniques. Because of the interaction between and the counteracting effects of the sympathetic and parasympathetic branches of the nervous system, self-paced tasks are difficult to analyze. It is necessary to have time marks, either recorded by the experimenter or based on other responses of the operator, to separate types of activities into meaningful categories. Otherwise, averaging will simply mask any effects of interest. (See Mulder and Mulder [1987] for further discussion.) Further, diagnosticity will be quite low if there is no way to relate the measured changes to ongoing behavioral activity.

Given that a technique is sensitive, additional procedures can be employed to improve diagnosticity. One such technique involves the use of time marks recorded during data collection. These marks can be based on external events, mission milestones, stimulus presentations, or operator responses. Having recorded some type of mark, the analysis can be locked on these marks. As will be apparent, such an approach to data collection and analysis can be critical for both sensitivity and diagnosticity.

Some analysis techniques, such as spectral analysis, require suitably long time segments to do the analysis. If the experimenter selects or samples a segment which is too short, the results may not be stable due to the small number of observations. By contrast, if too long an interval is selected, some of the effects may be masked. Here again, averaging and poor choice of the temporal scale may destroy potential sensitivity and diagnosticity of a physiological technique.

Physiological Background

Earlier, physiological techniques were referred to as a grab bag of techniques. To organize the techniques and to provide a benchmark for our evaluations and recommendations, we will provide a brief discussion of the physiology underlying the changes that various techniques are supposed to measure. This physiological overview will be referred to in the course of discussing workload techniques. It will also emphasize a clear rejection of the simplified arousal theory view that all physiological techniques were created equally (e.g., Hebb, 1955). To facilitate understanding for the reader, a schematic showing the relations of several physiological subsystems is shown in Figure 7-1.

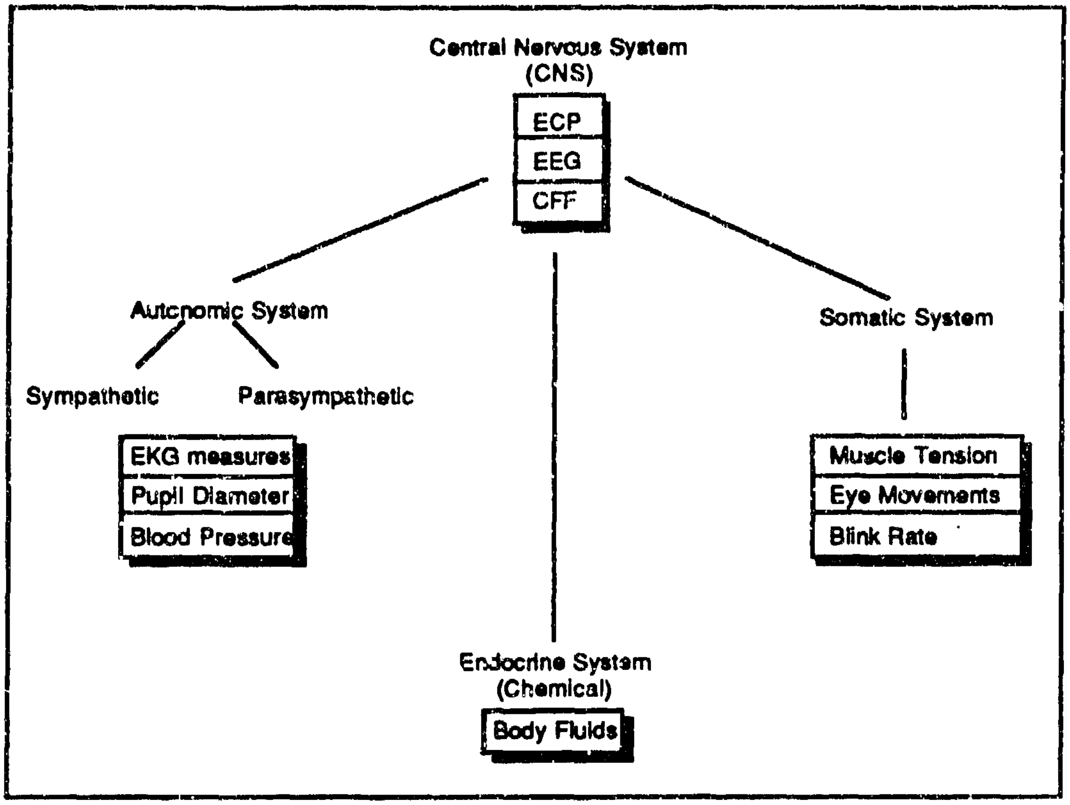


Figure 7-1. Illustration of the schematic relations among various physiologica: systems. The technique associated with each system is shown in a box.

The central nervous system (CNS) consists of the brain and the spinal cord. The CNS is actually composed of a number of distinct, identifiable neurological structures which are somewhat specialized to perform particular functions. Hence, superimposed on this structure are functional systems. For example, the language system is composed of a number of functional parts: hearing and analyzing speech, mediating or understanding the speech, a vocabulary, language rules and the organization and generation of speech, including not only ordering the words but control of the articulatory mechanisms to produce the sounds. The case studies of brain damaged patients demonstrate the existence and separability of these structures. There are a number of specialized cortical neurological structures which operate in unison to provide the capability of language with each structure contributing to the functional system.

The electroencephalogram (EEG) and evoked cortical potential (ECP) are techniques which reflect activity of the CNS and the cortex in particular. Electrodes placed on the scalp over particular neurological structures will measure very small changes in electrical potential occurring in the brain. Despite the use of identical recording procedures, the composition of the two techniques are quite different. The EEG contains a number of waves, some of which reflect general arousal. By contrast, the idea in the ECP is to average out these arousal waves and study components due to specific stimuli. To continue the language example, Chapman (1979) and his colleagues presented subjects with words selected along Osgood's semantic differential dimensions of Evaluative (good-bad), Potency (weak-strong) and Activity (fast-slow) while recording evoked potentials. He found quite different wave forms, implying differential brain functioning depending on the semantic dimension and affective meaning of the words. These results illustrate the level of detail that can be examined using physiological techniques.

Other techniques such as heart rate and pupil diameter are a function of the peripheral nervous system. The peripheral nervous system consists of all nervous cells outside the CNS including those entering and leaving the brain and spinal cord. The peripheral system is divided into two parts, the first is the **somatic** nervous system which includes sensory nerves from most receptors and motor nerves (effectors) for skeletal muscles. The second is the **autonomic** system which includes sensory and motor nerves serving the heart, glands, and smooth muscles. Eye movements are accomplished by three pairs of skeletal muscles (somatic system) while pupil dilation is under control of smooth muscles (autonomic system). Both of the these peripheral nervous subsystems are under general control of the CNS. It is the CNS and the autonomic system that are of principal concern for the measurement of workload.

The autonomic nervous system underlies emotional and motivational behavior. Any feedback system will have counter-acting influences and the autonomic system is no exception. It is divided into two parts which act in opposition to each other: The **sympathetic** and the **parasympathetic**. The sympathetic system activates the body and the parasympathetic serves to conserve the body. To illustrate, there is clear physiological evidence that stimulation of the sympathetic will result in heart rate increases and pupil dilation whereas stimulation of the parasympathetic causes decreases in heart rate and pupil constriction. Under normal operations, the two systems balance each other. However, an emergency may cause a brief imbalance which can have several different results depending on the timing of the two systems. Fainting, for example, is the result of reduced blood flow to the head caused by activation of the sympathetic system followed by a flood of activity from the parasympathetic.

Our brief discussion of some highlights of physiological function clearly shows the diversity of information available from the various rather specialized techniques available to measure physiological functions. Although one could say all measure CNS activity, most do not measure the activity directly. In the case of body fluids, the technique may be three or four steps removed from CNS events. There are also timing differences, neural activity is quick, chemistry much slower. Clearly the system is complicated with lots of antagonistic activity at all times. This implies that physiological techniques may reflect particular changes. However, if the technique does not show a change, one *cannot* infer high workload to be absent. Because of the rapidity of neural activity and the counterbalancing effects of the antagonistic systems, timing is of the essence. Failure to assess a physiological change associated with workload can reflect either an inappropriate technique or inappropriate data analysis.

Techniques Measuring Cardiac Responses

The heart is influenced by the autonomic nervous system and through this connection the heart is related to physical and emotional states. Heart rate is known to be related to the amount of physical activity (oxygen requirements), respiration, and thermal regulation. It can also be said that any factor which affects mental activities will also affect the heart. Thus, mental load and task demands will affect (and can be observed in) cardiac response. However, so can other normal factors such as the orienting response and the defense response; stressful and surprising events will also be evidenced. Additionally, factors such as age will result in changes in heart rate variability as well (Mulder & Mulder, 1987). Consequently, heart rate is a function of a number of forces which may be operating simultaneously.

Heart Rate

Many years ago, Darrow (1929) reviewed studies which seemed to show that looking at simple stimuli seemed to cause heart rate deceleration while stimuli that demanded cognitive processing were associated with acceleration. Since then, many studies have shown attention to the environment to be associated with heart rate deceleration. Acceleration is less clear, but certainly there is a relation between heart rate and the skeletal preparation for movement. Accordingly, unless the investigator is very careful to separate all of these influences, the increases and decreases may be masked. It seems modern investigators have had more difficulty with the technique; possibly the increased use of computer technology has moved the researcher away from the data.

There is some controversy with the OWL implications of heart rate (O'Donnell & Eggemeier, 1986; Wierwille, 1979). Not all investigators have found consistent results, or even results in the same direction. Since heart rate also increases with physical activity, one must take care when measuring mental workload that the technique is not contaminated by high physical activity conditions. Roscoe and Grieve (1986) and Wierwille and Connor (1983) have independently shown that the technique is sensitive to high stress/workload in which survival, embarrassment or similar emotions play a role. Similarly, long-term effects seem to be acknowledged. Sharit and Salvendy (1982) in discussing occupational stress, state, "The heart rate measure has undoubtedly been proven to be the most versatile measure of stress. p137." However, some investigators state that unless strong emotions are present, heart rate will not covary with workload (Hart, 1986a).

A recent report by Bauer, Goldstein, and Stern (1987) provides a departure in procedure from other studies and also illustrates one of the points made earlier about averaging. As indicated above, some investigators have failed to find consistent changes in heart rate as a function of task. Bauer et al. (1987), using the Sternberg task as a secondary task, collected a multiple set of measures that provides an opportunity to compare various measurement techniques. For data analysis on heart rate, they divided each trial into 18 time bins consisting of 950 ms each. In only eight out of the 18 time bins did task loading manipulation have a significant effect on heart rate. However, heart rate increased and then decreased as a function of time into the trial. These three intervals reflect different rule based activities. Averaging (which is done in analysis of variance) within just a six second interval to compare the cue, memory, and test intervals did not yield a significant difference for the three intervals. There was no effect of task loading but a clear effect of invoking different underlying processes. It is of note that their evoked cortical potential measure showed an effect of both task demands and task loading. While heart rate did not reflect task loading very clearly, it certainly showed clear differences related to what the subject was doing and when, that is, the changing task demands.

Heart Rate Variability (Sinus Arrythmia)

Heart rate variability (sinus arrythmia) is another workload measure relevant to heart rate data. It has proven to be equally controversial (O'Donnell & Eggemeier, 1986; Wierwille, 1979). Some of the inconsistency may be due to quite different analysis techniques; Kalsbeek ([1973], cited by O'Donnell & Eggemeier, 1986) has reported more than 30 techniques which have been used to determine variability.

Why look at variability? Simply on logical grounds one would expect an increase in heart rate to be associated with a decrease in variability; after all there is an upper limit on heart rate. As it turns out, there is a negative correlation (about -.40) between heart rate increases and heart rate variability. Even though there is a relation between the two, the fact that the correlation is modest indicates the two measures reflect somewhat different aspects of the physiological activity.

The spectral analysis of heart rate variability provides a method to separate out several frequency components stemming from different sources and seems to show promise as a measure. One peak, found around 0.35 Hz, represents respiration and a second peak reported at 0.20 Hz represents heart activity related to thermal aspects (Sayers, 1973). Some investigators suggest a thermal energy band from .02 to .06 Hz; arterial pressure from .07 to .14 Hz; and respiratory activity from .15 to .50 Hz (Aasman, Mulder, & Mulder, 1987). For our purposes the important peak, found around 0.10 Hz, is related to blood pressure and seems to be correlated with workload (Sayers, 1973).

The early work of Sayers has been followed by an increasing number of studies which show the .10 Hz component to be an effective indicator of mental activity, but care must be taken to factor out all confounding variables. Aasman et al. (1987) found a significant affect of task loading (2 or 4 items) in a continuous memory paradigm; the amplitude of the .10 Hz component decreased as the load on memory increased. These investigators attribute the change to the amount of effort expended, distinguishing between mental effort and mental workload (in our terminology different rules). Workload refers to dimensions of the perceived task demands and use of resources. Effort refers to what the subject is doing, the willingness to expend effort in the utilization of the resources. Overload, pushing the operator outside the workload envelope, however, results in a cessation of effort and a corresponding increase in the .10 Hz component.

Vincente, Thornton, and Moray (1987), in another recent study, used three levels of difficulty on a tracking task and had subjects give subjective ratings of difficulty, workload, and effort. Effort was defined as the amount of attentional demand; difficulty was defined as how hard the motor task was; and workload was defined as the overall level of demand on the task. These investigators did not find an effect of task difficulty on the .10 Hz component and only a marginal effect on the subjective estimates of effort. However, they found a correlation (.66) between the .10 Hz component and the subjects estimate of effort. Actually, seven out of eight subjects showed the correlation, the eighth did not. Of interest to the comments about data analysis, the tracking task is continuous. Timing of performance and the size of time samples used in the analysis are important (Mulder & Mulder, 1987).

Summary of Heart Measures

Both mean heart rate analysis and heart rate variability (spectral analysis) are based on measures of heart rate. Accordingly, one has a unique opportunity to extract two measures from a single technique. Mean heart rate, as indicated by the older literature and some recent studies, shows measurable changes as a function of task difficulty, quite possibly due to a generalized arousal component. Heart rate variability (.10 Hz component) often appears to be sensitive to task loading and fatigue (Egelund, 1982; Strasser, 1981). The majority of studies reported in the literature are looking for relatively subtle effects. In practical application, there may be some situations in which unknown but extreme demands may be made on the operator; heart measures would detect such situations.

Overall, one is not impressed with the consistency of the results using heart measures. This has lead O'Donnell and Eggemeier (1986) to conclude "For the present, therefore, heart rate and heart rate variability must be considered an attractive and promising but unvalidated measure of workload p 42-42." However, the European research groups (e.g., Mulder & Mulder, 1987; Strasser, 1981) have had reasonable success when the various confounding factors have been taken into account. Thus, in some applied situations, heart techniques may be appropriate.

Techniques for Measuring the Eye

Three separate visual structures are of interest in the context of OWL. These are

- The movements of the eye which are controlled by three pairs of muscles (horizontal, vertical and rotational movements) under control of the eye movement system,
- The pupil which is controlled through the autonomic system, and
- The lids of the eye which are under control of the somatic system.

Eye Movements and Scanning (Point of Regard)

Eye movements occupy a unique role in information acquisition. Because of the central role of vision and eye movements in information acquisition, many investigators have focused on information acquisition strategies reflected in eye scanning patterns to identify the source of information for decisions.

The goal of applied eye movement research has been to determine the scan patterns, how and where an operator gets information and in turn what he does with it. An assumption normally made is that dwell time (the length of a look) serves as an index of visual workload: The longer the dwell time, the more difficult to read the instrument. Current eye movement technology permits the investigator not only to monitor movement of the eyes but, with appropriate calibration, determine the point of regard, i.e., what was looked at.

In 1903, Dodge used film to record a reflected image of the eye which is still a useful technique. Since the time of Dodge, a number of techniques have been developed. (See O'Donnell and Eggemeier [1986] for a review of these various techniques and Hallett [1986] for a thorougn review of eye movement research.) While each of these techniques can serve a useful research function, few are useful in an applied context. Helmet mounted cameras filming the eye, much as Dodge did, have also proven to provide useful diagnostic information for OWL (Wilson, O'Donnell & Wilson, 1983).

Research shows workload can be predicted from changes in dwell times. These workload changes result from, for example, the difficulty of reading an instrument (Harris & Glover, 1984) or a change in mode of flying, autopilot or manual (Spady, 1978b). Waller (1976) showed eye movement data could be used to predict Cooper-Harper ratings, thus broadening the application of eye movement techniques for OWL estimation.

Diagnosticity of eye movement measures can be excellent. Wilson et al. (1983) were able to diagnose what a pilot was doing even when they could not obtain good evoked potential responses. The eye movement technique measures visual workload, but manipulations outside vision may increase general workload which can have an effect on dwell times.

Some eye movement devices can be expensive and require substantial data analysis capability, although many analysis techniques have been worked out (Harris, et al., 1955). Its potential as a workload analysis technique is highest among the techniques classified as physiological, however, it may not be practical at the present time for most Army applications. What can be of considerable use and much less costly, even though more obtrusive, is the helmet mounted camera – the Dodge technique.

Pupil Diameter - Pupil Dilation

It is well known that pupil diameter varies with a number of physiological and psychological variables. Beatty (1982) has reviewed the literature and concluded that the task-evoked pupil response reflects processing loads. In the context of our terminology, it appears to be sensitive to both rule changes and task loading. For example, pupil diameter changed both as a function of the phase of task (listen, pause, report) as well as the memory load (3 to 7 digits) (Beatty, 1982). Similarly, the measure has been shown to be sensitive to difficulty of tasks such as sentence comprehension and visual tasks involving comparison of letter pairs.

At present, because of the stringent restrictions on operator movement, the field application of pupil diameter measurement is minimal. To obtain good, accurate recordings, eye movement must be kept to a minimum; when the eye is at an oblique angle to the recording device, the two dimensional image of the pupil will be attenuated due to the geometry. Further, because the pupil varies with light levels independently of workload states, one must be careful to keep ambient light at a constant to avoid contamination of the data. It appears possible to remove both of those effects analytically, but this has not been done.

Blink Rate and Latency

Although blinking is subsumed under eye measures, the somatic motor pathway of the eyelids may be somewhat different from the motor pathway of the saccadic eye movement (Moses, 1970). There are two types of blinking, reflex and spontaneous. The spontaneous blink is of interest in the study of workload and can be conditioned. The duration of a full blink is about .3 to .4 seconds and occurs normally at the rate of about 2.8 seconds in men and 4.0 seconds in women. Blink rate of the eye has been measured using electromyography (EMG), sometimes called electro-oculomyography (EOG) when applied to eye research.

In the study discussed earlier using the Sternberg paradigm, Bauer et al. (1987) also measured blinking. Their analysis was parallel to that for heart rate and included three measures: blink rate, blink latency, and blink duration. For data analysis they divided each trial into 18 time bins consisting of 950 ms each. These bins were also blocked into intervals consisting of six bins each. Blink duration showed a decease over bins and an increase over intervals. Blinks occurred (blink latency) earlier following the cue than for other stimuli. For the blink rate analysis, the bin effect was significant; blink rate declined as an increasing function of time since the stimulus presentation. Of note, the blink rate declined from one every two seconds in the first bin to one every six seconds in the last bin. In eight out of the 18 time bins, the set size task loading had a significant effect on rate and overall set size had a statistically significant effect. Blink rate provides a measure directly related to task demands and to task loading.

Summary of Eye Techniques

Because vision is a major information acquisition sensory system, many investigators have focused efforts on determining how the system functions and acquires information under varying workload conditions. This has primarily focused on determining the point of gaze or look point of the eye. The three techniques considered in this section cover three different aspects of the nervous system. Of the three, the eye movement / point of gaze technique is probably the most important. The data derived from studies of eye movements have application in a number of workload situations: not only for instrument panels and computer displays but also visual search patterns used to detect events and targets. The cost and effort required to obtain and analyze eye movement data reduce the practical applicability of these techniques.

Pupil diameter has been shown to be sensitive to workload variations, especially the amount of mental load. However, measurement techniques do not lend themselves to field situations. These restrictions limit the technique to the laboratory. Blink rate is a technique that has received less attention, but it could be useful, especially in conjunction with other measures of eye behavior.

Techniques for Measurement of Brain Activity

Within the past 10 years, a considerable amount of effort has been devoted to identifying measures of brain activity that are reflective of underlying psychological processes that influence human information-processing and performance (Donchin, Ritter, & MacCallum, 1978; Hillyard & Kutas, 1983; Posner, 1978). Such efforts have offered promising results with respect to identifying brain activity patterns related to operator workload (e.g., Kramer et al., 1987). With respect to understanding human information processing and performance, researchers have recognized that measures of brain activity (e.g., cortical evoked potentials) are complex and their recording and analysis costly. Therefore, one is not likely to use these measures except when they provide data not easily available with more traditional behavioral measures (Duncan-Johnson & Donchin, 1982). A somewhat similar note of caution has been offered in regard to utilizing brain activity measures (e.g., cortical evoked potentials) as indices of mental workload (Kramer et al., 1987).

Electroencephalogram (EEG): Spectral Analysis

The electroencephalogram (EEG) is typically recorded from surface electrodes placed directly on the scalp. Such recordings can provide data on the brain's electrical activity during the performance of a task. Attempts have been made to quantify this electrical activity according to the predominant spectral frequencies that make up such brainwave activity. The premise is to identify those spectral frequency bands that are indicative of and reflect changes in workload. The EEG frequency bands that have received the most attention are 4-7 Hz (Theta), 8-12 Hz (Alpha), and 18-30 Hz (Beta).

In general, the findings support the conclusion that the percentage of low frequency EEG spectral bands (i.e., Alpha) increases during the course of prolonged and continuous performance (Parasuraman, 1984). Such findings have been interpreted as indicative of lowered arousal levels over time (Gale, 1977, O'Hanlon & Beatty, 1977). As a result, EEG spectral changes have been seen as reflecting general state changes within an individual (e.g., drowsy, alert). However, the relationship between these general state changes as shown by EEG spectral analysis and operator workload as indexed by performance changes is not always clear. For example, Gale, Davies and Smallbone (1977) used a simulated radar type task to show that subjects' performance declined as measured by reaction time (RT increased) during the course of prolonged performance which was accompanied by corresponding increases in the amount of the 7.5-9.5 Hz EEG spectral band (i.e., decrement in physiological arousal). In contrast, similar changes in the EEG have been reported by Fruhstofer and Bergstrom (1969) when subjects were relaxed and performed *no* task for a comparable period of time. The EEG spectral analysis approach has therefore been seen as indicative of organismic states which may or may not interact with or reflect workload (e.g., fatigue, boredom).

To illustrate this point, Howitt, Hay, Shergold and Ferros (1978) examined EEG changes for a single pilot during actual flights in a small two-engine transport aircraft. The flights were performed either as the first flight of the day, after a night of sleep deprivation, or after a series of daytime flights to assess sustained performance over the course of a workday. Each flight was considered to contain segments of differing levels of workload (e.g., single engine takeoff vs. maintaining steady level flight). Results showed a decrease in amplitude of EEG activity across several spectral bands (e.g., 8-12 Hz and 12-16 Hz) when sleep deprived flights or end of the day flights were compared to first of the day flights. These EEG changes were seen as reflecting organismic changes resulting from sleep deprivation (e.g., sleepiness) or prolonged work (e.g., fatigue). However, when comparisons were made to in-flight segments of different workload levels only the first day flights showed evidence in the EEG for reflecting workload levels (e.g., increased EEG amplitude for spectral bands) with concomitant increase in workload activity. By contrast, sleep deprived flights and end of day flights showed no signs in the EEG that were reflective of changes in workload levels during these flights.

Summary. EEG spectral analysis seems to offer means to assess changes in organismic states within an operator (e.g., fatigue, sleepiness) that may or may not show in performance. As a direct measure of workload, EEG spectral analysis is not an advantageous technique. Other researchers have voiced similar opinions (e.g., O'Donnell & Eggemeier, 1986).

Evoked Cortical Potentials (ECPs)

Usually, brain wave activity as measured by electroencephalography (EEG) reveals little in the way of discriminable patterns that can be attributed to operator workload. However, signal analysis techniques can be utilized to isolate specific brain wave patterns that are responses to external stimuli and may be used to reflect operator workload levels (e.g., Isreal, Wickens, Chesney & Donchin, 1980). These brain wave patterns found in response to external stimuli are called Evoked Cortical Potentials (ECPs) or Event-Related Potentials (ERPs).

The value of the ECP is based on the concept that brain waves reflect a combination of human sensory inputs (e.g., external stimuli/events) and cognitive processing (e.g., evaluating external stimuli/events). For example, when a stimulus is presented to the operator, a portion of the brain wave activity is a response associated with that stimulus. The remaining brain wave activity is considered as ongoing, unsynchronized, spontaneous activity that is not necessarily associated to the processing of such stimuli. By performing ensemble averaging across the time intervals following the multiple presentations of the stimulus, the ECP associated with such stimuli will be enhanced through this averaging while the spontaneous brain activity occurring in these time intervals will be cancelled out. Figure 7-2 depicts the relation between ongoing EEG activity, external auditory stimuli and signal analysis techniques used to extract the ECP associated with such stimuli.

ECP Components. As seen in Figure 7-2, the ECP is a complex wave form. It exhibits several components that are identified as either negative (N) or positive (P) peaks. In addition, these negative and positive components are further identified by their time course as measured from the external eliciting stimulus onset to their mean latency of occurrence (e.g., the P300 is a positive waveform component occurring at approximately 300 msec. after stimulus onset).

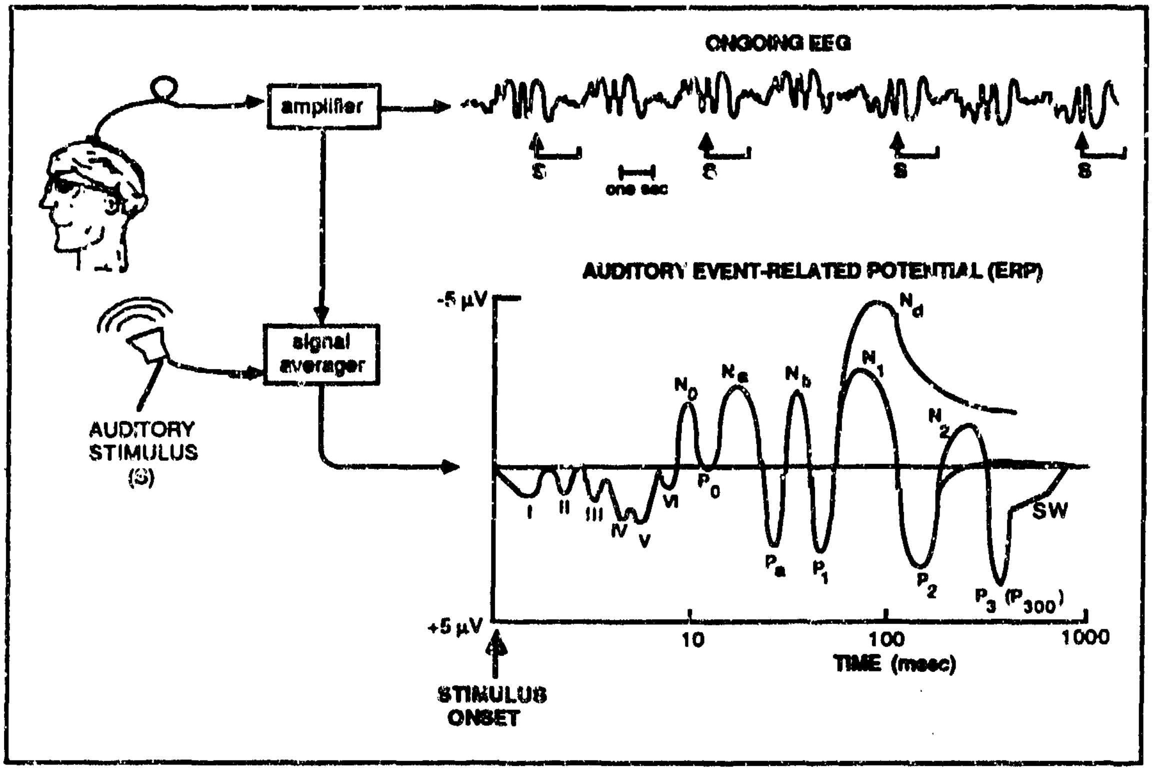


Figure 7-2. Depiction of the relations between EEG activity, external auditory stimuli, and signal analysis techniques used to extract the ERP associated with such stimuli. (Adapted from Hillyard & Kutas [1983]).

The early occurring components of the ECP, less than 250 ms from the onset of the external stimulus, have been characterized as being responsive to the physical nature of the external stimuli used to generate the ECP. For example, visual stimuli have elicited ECPs with identifiable early components that seem sensitive to manipulations of the physical parameters of such stimuli with respect to brightness (P200; Wastell & Kleinman, 1980), spatial orientation (N125; Harter, Previc, & Towle, 1979) and contour

(N150-235; Harter & Guido, 1980). Such ECP components are classified as exogenous (i.e., stimulus bound) since they are sensitive to the physical attributes of the stimuli (i.e., intensity, modality, and rate of presentation).

The later components of the ECP, those beyond 250 msec. from the onset of the external eliciting stimulus, are considered to reflect active cognitive processing of stimulus information. These ECP components seem to be sensitive to changes in the processing demands of the task imposed on the operator but not to changes in the physical characteristics of the eliciting external stimuli (Sutton, Braren, Zubin, & John, 1965). Those later occurring ECP components have been classified as endogenous components. The ECP endogenous component that has received the greatest attention is the positive waveform occurring approximately 300 msec. after the external eliciting stimulus onset (P300). The P300 has been examined as a measure to reflect cognitive processing activities as well as a measure to reflect workload levels. (See Pritchard, 1981 for a comprehensive review of the P300 literature.)

P300 and Cognitive Processing. The P300 waveform exhibits systematic changes in latency and amplitude that are used as evidence for its sensitivity to aspects of human information processing. In general, the P300 amplitude seems to be sensitive to the task relevance and the subjective probability of the eliciting external stimuli (Duncan-Johnson & Donchin, 1977). For example, the P300s elicited by task relevant stimuli are larger in amplitude than the P300s elicited by stimuli not relevant for the task to be performed (Roth, Ford, & Kopell, 1978).

The P300 latency appears sensitive to the time required to recognize and evaluate task relevant stimuli (Kutas, McCarthy & Donchin, 1977). That is, the P300 latency reflects stimulus evaluation time in the sense that identification and evaluation of a stimulus must be completed before the P300 is observed (Pritchard, 1981). This relationship between P300 latency and stimulus evaluation has been demonstrated to be independent of response selection and execution process. McCarthy and Donchin (1981) manipulated stimulus evaluation time by embedding a target word (P300 eliciting event) either in a matrix of # signs or within a confusable background of letters. Response selection was manipulated by changing the compatibility between the target word (right or left) and the responding hand. It was found that both visually distracting stimuli and stimulus-response incompatibility increased reaction time to the target words. Only the presence of the distracting stimulus backgrounds (letter backgrounds) had a significant effect on P300 latency (i.e., more evaluation time was needed to identify target words).

P300 and OWL. The findings just cited provide evidence that the P300 is sensitive to aspects of cognitive processing. Further, the relevance of P300 measures (amplitude and latency) to operator workload has been demonstrated in a series of studies conducted at the Cognitive Psychophysiology Laboratory at the University of Illinois. For example, Isreal, Wickens, Chesney and Donchin (1980) examined a display-monitoring task in which operators monitored 4 to 8 targets that moved across a television screen. Half of the targets were square-shaped objects and half were triangular-shaped objects. Operators were required to monitor one class of targets (squares or triangles) and to detect changes in either direction of movement or brightness. A secondary task was also required of operators. Operators were required to listen for high and low frequency tones that were presented during the performance of the display-monitoring task. They were instructed to count to themselves the number of times the high pitch tones (lower probability of occurrence than the low-pitch tones) were presented during the course of trial runs. They were told to report this number at the end of the trial-run. The P300 elicited to the rarer of the two auditory tones was used as a measure of operator workload levels. The concept behind such a measurement scheme is that the primary task will occupy operators' perceptual resources as a function of the primary task demands. More perceptual resources are needed to monitor 8 moving targets than 4 moving targets. As a result of this manipulation, the available perceptual resources needed to detect high frequency tones under high primary task demand will be less than under low primary task demand and therefore will be reflected in the P300s to such tones. The results of the study supported such a measurement scheme. The P300 elicited under control conditions (no primary task, counting of tones only) was highest in amplitude. This was followed next by the low perceptual demand condition (4 targets to monitor). Finally, the high perceptual demand condition (8 targets to monitor) was lowest in P300 amplitude. The conclusion to be drawn is that the P300 seems sensitive to perceptual task demands (i.e., workload).

The use of the relatively non-intrusive secondary task just described (auditory monitoring to detect infrequent occurring tones) has been called the oddball paradigm (Donchin, 1981). The oddball paradigm has been employed in several workload studies with similar results being reported. For example, Isreal, Chesney, Wickens, and Donchin (1980) and Wickens, Kramer, Vanasse, and Donchin (1983) have reported the P300 amplitude elicited by the secondary task (oddball paradigm) decreased as the perceptual task demands of the primary task increased.

Further evidence in support of P300 sensitivity to OWL has been reported by Wickens, et al. (1983). They were able to demonstrate with a primary tracking task in which discrete displacements of the tracking cursor served as the eliciting stimulus. The P300s associated with such a primary task, as contrasted with the secondary task, increased in amplitude as the perceptual demands of the task increased (i.e., operator workload). Kramer, Wickens and Donchin (1985) have reported similar results with P300s elicited by a primary tracking task.

The evidence presented in support of the P300 as a measure of OWL in this review has been confined to controlled laboratory situations. However, there have been attempts to record ECP with the use of the oddball paradigm in simulation type environments. Kramer et al. (1987) elicited P300s by means of the oddball paradigm while student pilots flew a series of instrument flight rule missions in a single-engine, fixed-based simulator. The flights varied in difficulty. The P300 amplitude discriminated between flights such that the more difficult flight mission elicited P300s lower in amplitude for the secondary task than the easier one. However, within-flight primary task demands were not distinguishable by the P300 amplitude; for example, takeoff, straight and level flight, holding pattern and landing. Natani and Gomer (1981) have reported similar success in using the oddball paradigm to elicit P300s with a low-fidelity flight simulation such that P300 amplitudes varied as a function of workload levels.

The P300 latency measure provides an accurate and reliable means to assess the time needed to identify and evaluate a stimulus prior to making a response. In addition, the P300 latency seems to be independent of response selection and execution processes (McCarthy & Donchin, 1981). As a result, the P300 latency can be used to determine the locus of performance changes that may occur. That is, if P300 latencies vary systematically with performance changes, one may conclude that identification and evaluation of stimuli are contributing significantly to performance changes such as increased reaction times. However, if P300 latencies remain invariant and stable to performance changes, such changes are not likely due to identification and evaluation processing. To illustrate, Gomer, Spicuzza and O'Donnell (1976) reported a study in which subjects performed the Sternberg memory-scanning task (Sternberg, 1969). Subjects were presented with probe letters of the alphabet (ECPs were elicited from these stimuli) and were asked to identify if the probes were members of a previously memorized positive set of letters. Memory load was manipulated by changing the number of letters in the memorized set. Both reaction time and P300 latency increased linearly as a function of memory set size for positive probe items. Such results support the inference that stimulus evaluation time (i.e., memory scanning) contributes greatly to reaction time scores in the Sternberg paradigm. In contrast, Duncan-Johnson and Kopell (1981) found that the Stroop effect (i.e., people respond slower to color words printed in a different color than the same color, e.g., blue printed in the color red as opposed to blue printed in the color blue) was mainly due to response incompatibility rather than perceptual interference (i.e., prolonged stimulus evaluation time). With the standard Stroop task, reaction time scores showed the usual interference between hue and word meaning. The P300 latencies elicited by such words however remained invariant.

Summary. The use of the Evoked Cortical Potential (e.g., P300) as an index of workload must be recognized as a highly specialized technique that requires a staff of highly trained personnel familiar with the recording techniques. There is also a need for expensive equipment and sophisticated software for the recording and analysis of the data generated. Beyond these considerations, there are other important technical as well as theoretical issues that may limit the applicability of using ECP as a measure of OWL:

- The ECP technique is based on producing an ECP in response to some time-locked repetitive stimulus event. Such eliciting ECP stimuli are usually controlled by the experimenter and are presented as secondary task stimuli (e.g., oddball paradigm). It is possible in some system applications (e.g., field testing and evaluation) such a stimulus would represent a form of intrusion and possible distraction to the operator. It may also not be possible to implement such a controlled type situation for some system applications.

- The ECP technique requires the use of electrodes and, in some cases, associated restraints are needed to reduce artifacts (e.g., eye movements that may contaminate visual evoked cortical potentials). As a result, the applicability of the ECP technique may be limited to controlled laboratory situations. To illustrate this point, Wilson et al. (1983) conducted a study with 12 A-10 tactical air command pilots. The study involved the implementation of various simulator emergencies conditions, whereby single evoked cortical potentials to auditory probe stimuli were recorded simultaneously with the occurrence of the simulated emergencies. Only three pilots' ECP data could be used out of 12 pilots. Artifacts in the EEG data of one pilot resulted in his rejection and the other pilots were rejected due to the fact that their ECPs failed to meet the ECP criteria for discriminability in order to be included in the data analysis. Such results point out the fragile nature of such recordings.

- ECP results may not show a strong relationship to other OWL measures. As a result, ECP data may be difficult to interpret with respect to their significance and implications toward system design decisions. For example, Biferno (1985) reported a study whereby subjects performed a compensatory tracking primary task and ECPs were elicited from auditory stimuli that were the call-signs designated for each participant. In addition, each subject filled out the NASA Bipolar scales to index subjective workload. The results were such that 4 out of 20 subjects exhibited significant correlations between P300 amplitudes elicited by their auditory call-sign and their weighted workload ratings. With only four significant correlations out of 20, the results are not encouraging with respect to a relationship between P300 amplitude and subjective workload ratings.

- Studies that have shown a relationship between ECP components (e.g., P300) and operator workload have been limited mostly to primary tasks that can be characterized as tracking type tasks. It therefore remains to be demonstrated that the ECP technique is applicable to other kinds of primary tasks that are now required of operators because of the advancement of technology (e.g., decision type tasks, data management and data fusion type tasks, and communications type tasks.)

Blood Pressure

Blood pressure reflects both cardiac output and vasomotor consequences of dilation and constriction of the blood vessels. (The vasomotor response serves two functions; to maintain body temperature and to direct blood flow to local areas.) The more blood pumped by the heart and the more the resistance the blood encounters in the vessels, the higher the blood pressure. Sympathetic activity tends to increase blood pressure by increasing heart rate and causing vasoconstriction.

Summary. Several studies have reported blood pressure changes with workload. Ettema (1969) showed relatively small effects over a short term but over a long term the pressure increased substantially. Similar results were reported by Ettema and Zielhaus (1971) who used auditory reaction time for the task. Nevertheless, the measure is not recommended for workload. One major delimiting factor is that this measurement requires the operator to sit still to get quality measurements. Further, blood pressure is a function of heart rate. One could eliminate a step in the physiological chain and measure heart rate directly.

Galvanic Skin Response (GSR): Skin Conductance - Skin Impedance

The galvanic skin response (GSR) is the measure of the resistance of the skin to the flow of electrical current. The resistance of the skin will change with degree of production of the sweat glands which are innervated by the sympathetic system. GSR is measured by applying a weak current through the skin and measuring the resistance. (Conductance can be obtained by taking the reciprocal of resistance.) Electrodes are usually placed on the palm or on the wrist. Skin potential is a related measure which is often used in modern research.

There is a large psychology literature employing the technique, however, not much has been done in the workload context. O'Donnell and Eggemeier (1986) do not even review the technique and Wierwille (1979) only discusses a few reports. For example, Kroese and Siddle (1983) studied workload while measuring GSR. They varied the stimulus presentation rate of digits; the task was to pick out odd and even sequences. GSR was measured on irrelevant tones presented during the task. They showed skin conductance to decline (habituate) more slowly for higher workload conditions (faster stimulus rates). The fact that the GSR habituates with repeated presentation of stimuli, makes it less suitable for workload research and evaluation than many other techniques. The habituation in amplitude of the response may vary with workload but it has to be measured over a series of presentations and then a new, novel stimulus must be presented. It might be useful for perceived emergency conditions.

Summary. GSR has been shown to be related to short-term general arousal effects. Sensitivity is reasonable; diagnosticity is low. For both theoretical and practical reasons, it is not recommended as a preferred technique in OWL assessment.

Electromyography (EMG) (Muscle Potential)

General arousal theory would claim that an increase in mental activity would be accompanied by an increase in muscle tension. Electromyography is used to provide a measure of muscle tension and activity. It is, however, a measure of somatic rather than autonomic nervous system activity and because of this it is a rather indirect measure of workload.

Muscular tension is related to both physical and mental activity. Indeed, dealing with inappropriate muscle tension is one of the more common approaches to athletic psychology (Nideffer, 1976). In tennis, for example, missing the first serve may cause the player to 'tense up' with the result that the muscles are tighter and the loss on the second serve is not as high. The consequence often is that the second serve is also missed. Clearly, mental activity has caused a change in muscle state.

The electrical potential created by motor units of the muscle reflects both the force exerted by the muscle and the tension in the muscle. This can be measured by implanting electrodes in the muscle or, more feasibly, by measuring the surface potential. In physical work, it is believed there is essentially a linear relation between muscle activity and the recorded potential. This permits the measurement of both (a) immediate work (forces exerted) and (b) long-term activity. In the former case, the absolute forces required to move or operate can be measured. In the latter case, temporal analysis of speed and degree of shift will show different spectral characteristics.

Summary. There appears to have been little research using this technique in the last ten years. Wierwille (1979) reviewed a few studies which show increased tension to be correlated with increased workload; O'Donnell and Eggemeier (1986) reviewed the same studies and came to a similar conclusion. Although the technique reflects workload changes, it is a technique that measures the somatic system and is only secondarily tuned into mental workload. There are also more practical ways of measuring physical activity such as video taping movements and analyzing them later.

Critical Flicker Frequency (CFF)

CFF is that transition frequency at which a flickering light passes into perceived steady state, fusion. A tremendous amount of research has gone into this phenomenon over the last century. Brown (1965) has reviewed much of the work on intermittent stimulation up to the date of his review. (See Watson [1986] for a thorough discussion of this approach as well as a current review of temporal sensitivity.] The relative importance of the phenomenon for this report is, of course, the application of the technique for measuring workload.

CFF is a diffuse but direct measure of CNS functioning. CFF occurs at frequencies between 50 and 70 Hz depending on contrast and illumination (Brown, 1965; Watson, 1986). Because cells below cortical level have been shown to have capabilities to respond to stimuli at much higher frequencies than behavioral CFF, it has been taken to be an index of physiological functioning of the cortex. Further specification is provided in a study by Wilson and O'Donnell (1986) which isolates several aspects of CFF with respect to physiological functioning. The procedures used are too complicated and specialized for applied work; nevertheless, the results are related to the diagnosticity of CFF. They used steady state evoked potential to separate out three frequency ranges of flickering stimuli. These ranges are centered at approximately 10 Hz (low), 18 Hz, and 50 Hz (high), each with differing amplitudes of the averaged signal, following from the work of Regan (1977). These results indicate that high frequency transmission is related to sensory-motor portions of the scanning task while the medium frequency is related to cognitive portions. This work is suggestive that CFF changes are related to sensory functions of the CNS.

Wierwille (1979) reviews one study which suggests CFF changes are related to fatigue but not cognitive workload in any direct or consistent manner. Oshima (1981) has summarized his work on CFF as a measure of mental fatigue. Most of this work is in Japanese and therefore, procedural details are not readily available. However, he suggests CFF is an effective technique to measure fatigue. He also shows substantial variation as a function of diurnal rhythm. Brown (1965) also reports effects of diurnal rhythm in his review. Fatigue, anoxia, effects of drugs, state of arousal, and age are among other factors shown to influence the CFF (Brown, 1965). Wilson and O'Donnell (1986) have shown a unique and a high degree of stability of response to flicker for several individuals over several years.

Summary. The CFF technique can be applied in a short period of time. In general, psychophysical measurement tends to be quite reliable and stable when extraneous factors are controlled. However, care must be taken with the technique to evaluate all of the factors which have been shown to influence CFF. Changes in CFF can be due to a number of variables, but when these are factored out, it appears to be a broad index of the efficiency of CNS functioning, especially the sensory component. It could be used effectively to evaluate long term effects of workload during sustained operations and the depletion of resources.

Body Fluid Analysis

Body fluid analysis is one of the few techniques available for the assessment of sustained or long-term effects of workload. Three body fluids are known to change their chemical composition as a function of long-term workload and stress: Blood, urine, and saliva. Recent work has concentrated on urine and saliva because these two can be obtained relatively more easily than blood samples. Periodic urine collections may be difficult to accomplish because of requirements to produce on demand. Both urine and salivary fluids may be particularly difficult to obtain just after intense stress. Indeed, the psychoendocrine approach has been adopted by many researchers in stress research. Sharit and Salvendy (1982) provide a summary of both theoretical and empirical work using the approach.

The compounds typically assayed involve both sympathatic nervous system and bodily metabolic functions. According to Wierwille (1979), the concentrations of compounds in the urine or parotid fluid that are examined and their indications are

- norepinephrine - sympathetic nervous system activity,
- epinephrine - sympathetic nervous system and adrenomedullary activity,
- 17-hydroxycorticosteroid (17-OCHS) - adrenocortical activity,
- urea - protein metabolism,
- sodium - mineral metabolism,
- potassium - mineral metabolism, and
- sodium to potassium ratio - metabolic balance.

The usual procedure is to gather samples before, during and after a prolonged task. The samples are then analyzed chemically for concentrations of compounds suspected to be related to high workload. Timing of the collection of the fluids may be critical when measuring the sympathetically induced changes. This timing issue is less critical for physical activity and the metabolic measures. The technique is believed to be sensitive to prolonged stress and strain. It is also likely to be sensitive to physical workload, particularly for compounds associated with sodium, potassium, and urea. The technique is useful for assessing possible long-term effects but is not recommended for short-term effects (Wierwille, 1979).

An alternative to using body fluid analysis might be to use a subjective method, in particular a mood scale. Frequently used in stress research, the mood scale offers a reasonable alternative to the chemical assay method, reduces the resource requirements, and can be administered relatively quickly.

Overall Summary

Physiological techniques assess a variety of physiological subsystems which are directly or indirectly influenced by workload variations. Some of these techniques are highly specialized to examination particular parts of the system during high workload. Because of the rapidity of nervous system activity and the counterbalancing effects of antagonistic systems, the timing of measurements is critical. Every technique reviewed has been shown to be sensitive to workload and almost every technique has been shown to have failures. One of the important aspects involved when applying any physiological technique is the recognition that various subsystems operate in opposition. Accordingly, data analysis plays an important role in the success or failure of OWL assessment for many of the techniques.

A number of physiological techniques have been used in the evaluation of workload. The discussion can be summarized into four broad categories:

- **Heart.**

Mean rate. Heart rate has been snown to be sensitive to workload variations. The technique is controversial, but certainly will reflect high stress/workload

Variability (sinus arrythmia). Also controversial and a technique which requires care in data analysis, heart rate variability has been shown to be sensitive to workload.

- **Eye.**

Eye movement measurement is the most promising physiological technique in the applied context. Much of the basis for usefulness of the technique rests on the high degree of reliance on visual information in modern systems. Dwell times give an indication of importance and/or the difficulty of interpreting an instrument or display. The technique does, however, require considerable resources. As indicated by a considerable body of research, eye movement techniques are certainly sensitive and have a capability for diagnostic information for OWL (Harris, Glover, & Spady, 1986).

Pupil dilation has been shown to be sensitive to workload variations, however, restrictions required to obtain clean measurement limit the field application of the technique.

Blink rate and associated measures such as latency have been shown to be sensitive to workload vanations.

- **EEG/ECP.**

These two techniques measure electrical activity of the brain. While they have been used quite successfully in the laboratory to assess cognitive states and their relation to OWL, the techniques require considerable resources and can be difficult to implement in field situations.

- **Other Techniques.**

Blood pressure. This measure is not recommended because of the confounding of cardiac output and temperature regulation.

Galvanic skin response (GSR). This has been shown to be sensitive to mental load, however, the effect is one of slower habituation. This tends to less useful as a workload technique.

Electromyography (EMG). Increased muscle tension may be an immediate consequence of increased workload, but not necessarily. For the purpose of measuring longer term physical work, the technique would be useful.

Critical flicker frequency (CFF). CFF can be measured easily and reliably. It appears to be sensitive to longer term effects, especially for the sensory system.

Body fluid analysis. This is a general technique which can be used to detect long term effects of workload and stress.

CHAPTER 8. MATCHING MODEL

The purpose of the matching model is to assist the user in selecting OWL measures for the Army system to be analyzed. The goal is to use all of the information available in the best way possible to match the requirements of the user with characteristics of the OWL techniques. The analysis of interest to the user may be for an Army system going through the traditional materiel acquisition process (MAP), or through Army Streamlined Acquisition Process (ASAP), Product Improvement Program (PIP), PrePlanned Product Improvement (P3I), or Non-Developmental Item (NDI) procurement. One reason for the Matching Model is the complicated nature of the OWL measure selection process. Another important reason for the matching model is to take into account the needs and requirements of the user, and the intended application of the results.

It has been suggested that the Army does not have sufficient human factors personnel available to deal with any but the most pressing operator workload issues. This was partly revealed in Army interviews (Hill, Lysaght et al., 1987). Further, with the emergence of MANPRINT, there is an even greater need and demand for human factors analysis in general and OWL analysis in particular. Clearly there is need for more expertise and greater distribution of OWL information within the Army community. The question then is how to provide such expertise within existing frameworks and organizational structure. While there are a number of alternative solutions such as bringing in more experts, by far the best alternative (and least expensive) is to use a computerized Expert System approach. An Expert System, for present purposes, is a method of formalizing the considerations involved in selecting OWL measures to apply to analysis of Army systems in various stages of development.

When one calls in an expert, one expects to get answers to the problem at hand. No answers are possible, however, without clearly stated questions. Hence, the expert will often begin by asking a host of questions, starting with very general issues and gradually asking about more and more detail, finally coming up with one or more suggestions. The thought processes generally follow a relatively consistent line whether the expert be Sherlock Holmes solving a mystery, Einstein developing relativity theory, or a practitioner developing a line of analysis for measuring OWL. Although not always formalized, the steps are: first, develop a system model which organizes the available facts; second, determine what pieces are missing and where the gaps are; develop the hypotheses; and third, generate specific questions to be answered. The point we wish to make is there is nothing so practical as having a system model. This system description or model provides an organization for the operator behaviors involved and a framework which is extremely helpful in posing the questions. Such a model can often be obtained from analytical techniques; analytical techniques often have an important secondary function since they provide the initial basis of an Army system model of the system which facilitates the generation of questions and subsequent answers. In the next section, we will begin to formalize the steps a human factors workload expert would follow in selecting an appropriate battery of techniques.

There are a variety of analytical techniques which can be used during early concept phases and also later in development. Not all of these analytical procedures have been fully validated. However, in order to be validated, they have to be used. Accordingly, we will suggest techniques that appear to be appropriate, independent of validation. In our discussion, we will describe narratively and show graphically the reasoning underlying the selection of techniques from the analytical category of the OWL technique taxonomy. Then we will consider some examples and case studies for empirical techniques. Following that, we will lay out the considerations for an overall, general matching model which includes both analytical and empirical techniques. During our data collection, the Army community expressed a desire for a computer-based rather than a written manual (Hill, Lysaght et al., 1987). To respond to this desire our expert system will build on developments incorporated in W C FIELDE (Workload Consultant for FIELD Evaluation) which was built to deal with empirical techniques in an aviation context (Casper, Shively, & Hart, 1987). At the end of this chapter, we will provide some background on computerized expert systems.

Analytical Matching Model

Analytical workload assessment techniques can and should be utilized throughout a system's development cycle, but are especially important at early, pre-hardware stages. As suggested in Chapter 3, there are few good predictive techniques and many of the analytical techniques have limitations. Nevertheless, the tremendous cost / benefit value of recognizing and diagnosing problems early on makes the use of these techniques imperative.

This section describes the core of the analytical methods segment of the overall matching model. It will assist the workload analyst to make intelligent decisions as to which analytical methods to use for a specific situation. First, the reasoning underlying the model is explicated in narrative form. This presentation is high-level and is intended to be exemplary, not comprehensive. Then the reasoning is formalized in logic flow graphical descriptions. We have chosen to begin the formalization process immediately rather than wait until after validation; in this way, creation of the matching model in the form of an expert system is facilitated.

System Considerations

The logic underlying this first-cut analytical component is explicated in the following system considerations. Hopefully, as a result of this report and others (Hill, Plamondon, Wierwille, Lysaght, Dick, & Bittner, 1987) analytical techniques will receive a boost toward more development and validation. The main considerations for analytical procedures are:

- What is the stage of development of the system?
  - If the system exists only on paper, then analytical techniques are the techniques of choice.
  - Otherwise, if some hardware exists, then both analytical and empirical techniques are possible. Please note, however, that one should not utilize empirical techniques without a very clear picture of the questions to be answered.
- Has a mission scenario been developed for the system?
  - If the answer is no, then one must be developed. It is absolutely essential to have a definition of not only what the system must accomplish but also specification of the accuracy required and the available time in which it has to be done. Additionally, the conditions under which the scenario is to be accomplished should be specified. The scenario becomes the specific framework within which OWL can be assessed, and time and accuracy become the measures of effectiveness (MOEs) within which the man/machine performance must fall.
  - If the answer is yes, then one can proceed.
- Has any workload analysis been done on similar systems?
  - If the answer is no, then we start fresh doing an overall analysis, probably in terms of task analysis or simulation.
  - If the answer is yes, then one should build on the analysis which is already available. Certainly, one would want to compare the new system with other existing systems via Comparison Analysis.
- Has any workload analysis been done on this system?
  - If the answer is no, then we can skip this question.
  - If the answer is yes. then presumably more detailed questions should be addressed. It may then be appropriate to analyze a specific portion of the system in detail using one of the mathematical model techniques or operator simulation.

Real world constraints

Having identified system issues, one also needs to consider real world constraints imposed on OWL analysis. These constraints include limits on the time available to do the analysis (How fast must the analysis be done? In what time frame?), the manpower available to do the analysis, the level of expertise of the staff available (which will have an impact on the length of time required and how much can be done) as well as the level of detail required in the analysis. Additional constraints may exist in the form of computer facilities to run simulations, both on the hardware side and the software side. (However, it should be pointed out that both Micro SAINT and HOS-IV are available to Army users.) Applying these constraints may lead to ruling out certain types of analysis techniques. For example, if only two weeks are available, then one might only use expert operator opinion to identify chokepoints. Otherwise, a more detailed analysis should be done.

Decision logic

The flow of the decision logic is illustrated in Figure 8-1 and elucidated in the following outline. This figure does not contain all of the appropriate detail but serves to show the principal steps, primarily for systems in the PreConcept or Concept Exploration Stages. However, analysis of workload is an iterative process and these techniques will be useful at any point in the analysis process. Feasibility checks, shown in the upper right of the figure, are also repetitive; the proposed analysis must be compared against real world constraints at various steps in the process. Specifically, the feasibility issues are

- Time constraints, how much time is available to do analysis?
- Manpower constraints - How much manpower is available to do analysis?
- What is the detail required in the analysis?
- What is the required accuracy of the analysis?
- Facilities - are computers and software available for simulation?

Step I: Has any OWL analysis been done on this system?

Alternatives:

If no, proceed to Step II.

If yes, proceed to Step IV.

Step II: Are any relevant data available? Check the MANPRINT ON-LINE database in the Soldier Support Center for possible databases which may contain relevant information. Also check the Manpower and Training Research Information System (MATRIS) office of the Defense Technical Information Center, San Diego, for material from their MANPRINT database. The Army Research Institute (ARI) and the Human Engineering Laboratory (HEL) have strong human factors engineering expertise and it would be well worth while contacting one or more individuals in these organizations. If no relevant information is found, go to Step III; otherwise if relevant information is found go to Step IV.

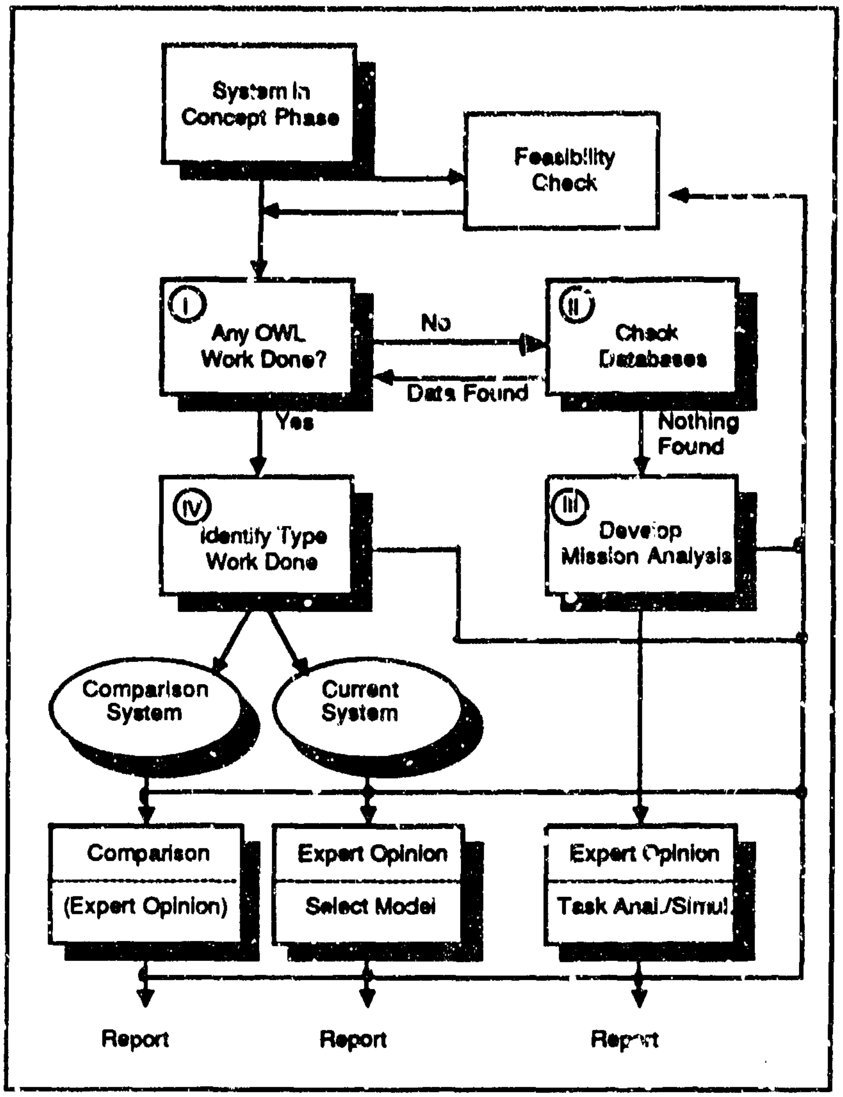


Figure 8-1. Diagram of OWL analytical Matching Model.

Step III: Has the mission scenario been developed?

Requirement: A mission scenario. If not available, it must be developed before proceeding. (A feasibility check should also be done at this point.)

DO

Expert opinion Use expert opinion from one or more individuals to identify questions of interest. Experts should be able to identify possible chokepoints to focus on in the analysis.

AND

Perform task analysis

OR

Simulation.

THEN

When hardware arrives, other techniques, especially empirical ones, can be used.

Step IV: Is the previous OWL analysis information of interest for a comparable system or on the system ?

If OWL research has been done on a comparable system

THEN DO

Comparison Analysis between the older system and the current system.

OTHERWISE select one of the following specific issues for the current system.

Issue 1: Re-evaluation or additional work needed, that is, inadequate information is available.

DO

Expert opinion Use expert opinion from one or more individuals to identify questions of interest. Experts should be able to identify possible chokepoints to focus on in the analysis.

AND

Task Analysis

OR

Simulation.

Issue 2: Functional re-allocation of man and machine tasks.

DO

Expert opinion Use expert opinion from one or more individuals to identify questions of interest. Experts should be able to identify possible chokepoints to focus on in the analysis.

AND

Simulation

Issue 3: Specific design issues (clarify data and issues).

DO

Expert opinion Use expert opinion from one or more individuals to identify questions of interest. Experts should be able to identify possible chokepoints to focus on in the analysis.

AND one or more of the following

Math models:

Anthropometric model

Sensory model

Manual Control model

Queuing Theory model

Task analysis:

Cognitive task analysis

Simulation:

Detailed network models or HOS may really be the only simulation models specific enough to analyze design issues.

Performance model - Card, Moran, and Newell (1986).

Empirical techniques:

Part-task analysis can also be accomplished with empirical techniques.

One of the techniques recommended throughout is the elicitation of operator expert opinion. Often the individual developing the OWL analysis does not have direct, first hand experience on the operation of the system. Use of operator experts can both save time and provide a focus on operator chokepoints. As one can imagine, the definition of an expert varies widely and one needs to be aware of the background of the expert. For example, an airline pilot is certainly an expert on aviation, but would not normally have substantial background on advanced avionics or advanced display technology. A test pilot, by contrast, would likely have a much richer and broader experience with new devices and technology and would be able to identify more quickly the potential trouble spots. This does not mean that expert opinion is not useful, it simply means it should be put in the context of the experience of the operator.

Task analysis is also a technique utilized very generally. Typically, a task analysis is done in concert with, or directly following, a mission scenario development, which is required for all systems. The task analysis forms the basis for performing more formal analytical techniques such as mathematical modeling and simulation, and serves as a guideline for any empirical work.

Empirical Case Studies

Portions of the empirical matching model are already available in the NASA Ames Expert System W C FIELDE (Casper et al., 1987). This system has been reviewed by experts in workload research and has gone through several revisions. The Matching Model outlined in this chapter is anchored on the structure of W C FIELDE. However, the workload approach as characterized by W C FIELDE omits some issues of major interest to the Army community. W C FIELDE, in particular, does not have the capability for direct comparison of two or more systems nor does it consider individual differences. Of potentially more importance, it also does not consider conditions under which the system must operate, such as battlefield conditions, or system support requirements. Many of these conditions cannot be tested except in an analytical way. These are not criticisms of W C FIELDE. Most of the OWL literature, being more academically based, frequently addresses issues directed to a theoretical interest, instead of those central to the goal of application. These academic researchers by and large have not only ignored some issues, but have actively sought to reduce or eliminate them as contaminates of the 'real' issues they wished to study. Although theoretical research is productive and important, it is not sufficient. Individual differences, which are frequently controlled experimental factors in research laboratories, and the comparison of combat systems are extremely critical factors in Army systems development.

It is our intention to develop a complete and integrated matching model for both analytical and empirical techniques, and one which will provide OWL measures sensitive to individual differences as well as to comparing several systems. In this section, we will discuss some examples of system design issues and provide recommendations for selecting empirical measures and appropriate analytical techniques for situations of immediate interest to the Army. The issues are focused at different system evaluation problems. For example, sometimes a workload study may be devised, other times the study has already been done. The measures suggested are the minimum one should collect. The cost of collecting the data and the analysis requirements have been taken into account in our recommendations.

System Design and Development Example 1

Description of Example. You have a system which requires that the operator routinely performs several tasks or sub-tasks (e.g., tracking targets, radio communications, weapon delivery, etc.) in order to carry out a mission. You are interested in knowing whether an operator can adequately handle the system. Specifically, what are the limits in the operator's performance before the operator's performance deteriorates, that is, show signs of overload? The following steps are recommended and are also illustrated in Figure 8-2. (The numbering of the steps matches the Roman numerals in the figure.)

Step I: Identify the conditions under which this system will be used. Then, identify those conditions which can be tested. A feasibility check is appropriate at this point in the process.

Step II: Define your measures for the primary task, including the overall system measures (Type 1) and operator response measures (Type 2) as discussed in Chapter 4. These performance measures may yield important information on overload, system instability, as well as permitting inferences on performance rule changes. In addition, consider the use of SWAT or TLX to get quantified measures about the operators opinions about workload as well as interviews of the operator to get additional detail. Finally, the heart rate measure can be useful as a physiological index and can yield some additional information. Depending on the context, one might wish to consider use of a helmet mounted eye tracking camera for diagnosis. Video taping the operator during performance of system tasks is highly recommended and can be used for delayed, retrospective TLX (or SWAT) ratings.

Step III: Perform the study. But before commencing, review the feasibility. Are the techniques feasible? What are the time constraints? How much time is available to do analysis? How much manpower is available to do analysis? What is the detail required in analysis? What is the required accuracy of the analysis? Are computers and software facilities available?

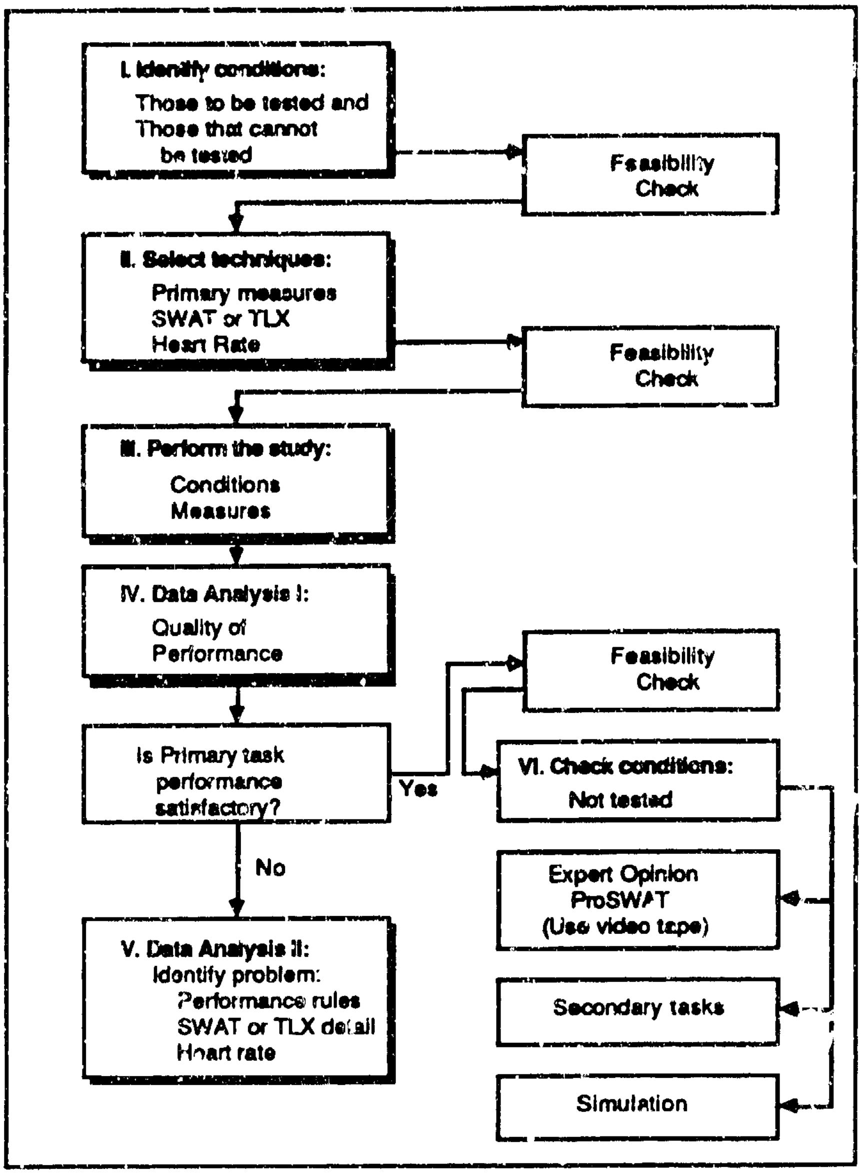


Figure 8-2. Illustration of the matching logic for determining the selection of OWL techniques for assessing overload in a system.

Step IV: Check the primary measures for performance decrements and OWL problems. If there are indications of problems, proceed to Step V. If there are no apparent problems, jump to Step VI.

Step V: Examine the fine structure of the primary task to look for performance rules. Examine the detail of the subjective scales (TLX or SWAT) to try to diagnose and identify the specific issue or problem. The heart rate data can be analyzed to give more detail and temporal locus of the problem.

Step VI: If there are no apparent OWL problems more work may still be required. There may be a need to look at the conditions which were not tested, such as environmental extremes. This could be done by a mixture of analytical and empirical techniques. The analytical portion would include use of expert operator opinion both through interviews and quantification through the use of ProSWAT (or ProTLX). (If video tapes were made originally, the video tape may be useful here for replay to the operator for retrospective ratings.) Model simulations of the potential chokepoints could also be done, such as Micro SAINT or HOS to test extreme conditions. The empirical techniques would focus on secondary tasks in the attempt to drive the operator to higher workload levels. The secondary task results coupled with the primary measures will yield important data about strategies and identify borderline workload portions.

System Design and Development Example 2

Description of Example. You have two alternative designs of a system or sub-system which have been shown by previous testing to be essentially the same (no differences) with respect to primary task measures. In this situation, you are faced with what appears to be two comparable designs. Which design do you choose? Since the testing is already done, this can result in some serious problems as will become apparent in the discussion.

Step I: Identify the conditions under which this system will be used. Then identify those conditions which have been tested. A feasibility check is appropriate at this point in the process.

Step II: Determine the level of data available. The data one would want are those described in Example 1; specifically, complete primary task data and the subjective scale data. Additional data are always welcome, especially video tape of the operators. If the primary task and subjective scale data are available, go to Figure 8-2 and follow the accompanying description, especially from Step IV on. If these data are not available, there are a few things one can do.

- Redo the OWL analysis as described in Example 1, Step IV
- If video tapes are available, then one can ask operators to use SWAT (or TLX) retrospectivelv on the video tapes.
- Use analytical techniques as described in the Analytical Matching Model.
- If no data are available and you cannot do any of the above, our advice is: **Don't ever get into this situation.** All you can do is start at Step I as described in Example 1.

System Design and Development Example 3

Description of Example. You have a system that is under a Product Improvement Program (PIP) or P3I for enhancements or modifications. You are interested in whether the operator can handle the new capabilities and/or new functionality that is planned.

Step I: Identify the conditions under which this system will be used. Then, identify those conditions which have been tested. The system may be manageable under test conditions but may not be manageable under more extreme, e.g., combat, conditions. Note the application of a feasibility check at this point in the process.

Step IIa: Plan a task analysis to determine if the new system capabilities involve new tasks which are added on, or if the new system capabilities will help the operator perform his duties, or both.

Step IIb: Plan a comparison analysis incorporating expert opinion.

Step III: Since this is an existing system, a workload study can be conducted on the present system with a secondary task. This task could be to measure the operator's spare capacity and to look for performance changes and especially performance rule differences. The secondary task should be selected to be comparable if not analogous to the planned modifications.

Step IV: Do a feasibility check before starting. Are all the techniques feasible? How much time is available to do analysis? How much manpower is available to do analysis? What is the detail required in analysis? What is the required accuracy of the analysis? Are computers and software available for simulation? Are computers available for data analysis?

System Design and Development Example 4

Description of example. You have a system under test and evaluation. You are not only interested in knowing whether the system can be handled by operators within the context of a mission scenario but also where the high workload areas are that could lead to operator workload problems.

Step I: Identify the conditions under which this system will be used. Then, identify those conditions which have been previously tested. Again, do a feasibility check at this point in the process.

Step II: Plot the mission profile. An example of this in aviation would be take off, ascent, cruise, descent, approach, and landing. Use expert opinion to determine those mission segments which have higher workload than others. If you wish, use ProSWAT (or ProTLX) to quantify the expert opinion.

Step III: Do a cognitive task analysis to identify the goals and strategies. This will assist in selecting primary measures and secondary tasks.

Step IV: Then perform the study as suggested in Example 1.

Step V: Are the analytical techniques feasible? Are the empirical techniques feasible? How much time is available to do analysis? How much manpower is available to do OWL analysis? What is the detail required in analysis? What is the required accuracy of the analysis? Are computers and software available for simulation? For data analysis?

System Design and Development Example 5

Description of Example. How would you deal with individual differences, personnel considerations and OWL while developing and testing a system?

This example on individual differences often falls in the cracks between operator workload and personnel issues. The Army has a wide range of personnel capabilities and any OWL analysis should include this consideration. The operators performing test evaluations may not be representative of the overall population. Whether the operators are representative is something that can be evaluated.

There are a number of tools being created for the concept/design phase of system development which will help to answer personnel/OWL problems. In particular, the MANPRINT Methods Product 6 is an analytical tool which is being designed to address the questions of what kinds of personnel characteristics are necessary to operate (and maintain) systems.

At present, the evaluator can do some straightforward things such as get as much information from the test operator's personnel file as possible. These items would include the ASVAB, any MOS information available, and Skills Qualifying Test (SQT). It is not suggested here that any one or all of these items together will provide detailed predictions of performance on a system. They will permit some relative comparisons of the general capabilities of the operators tested with the pool of operators for which it was designed.

The Basis of a General Matching Model

The objective in developing a general matching model is to provide a basis for the systematic selection of a good, if not optimal, set of workload assessment techniques for a given circumstance. As this is an ambitious undertaking, we have begun what is clearly an evolutionary process. Such a beginning will serve to stimulate rapid growth in the area of workload assessment. The model offered here builds and expands upon the concepts contained in W C FIELDE.

The particular use of the matching model will depend on the situation of the user. The user may be the OWL assessor, or the user may designate that role to a designer, an engineer, etc. Many variables properly affect the selection of an assessment technique battery so all appropriate personnel involved in the development and evaluation should collaborate in this decision. The model offered is designed to enhance this collaboration.

Specific Goals and Cojectives

The specific goals of this effort were to develop the framework for a user-centered expert system / decision aid technique using:

- **The Matching Model** to guide the user to the appropriate workload methodology;
- **The OWL Information System** to guide the user to the appropriate background literature;
- **Other Databases** to guide the user to the appropriate and available comparison systems; and
- **Other tools** as may be available or are in development such as MANPRINT Methods being developed by ARI.

The interrelation of the component parts of this approach is illustrated in Figure 8-3. In developing this overall model, our guiding principle was and is that the user need not be an expert in human performance technology, statistics and data analysis, laboratory work, and using computers. It was assumed that the user will be responsible for deciding on workload analysis techniques and will be responsible for getting the analysis done. Furthermore, it is anticipated that the user may not be totally familiar with the Army system acquisition processes. The emphasis is not to make a user into a human factors engineer but the goal is to make the user more knowledgeable about what needs to be done and what are the available and preferable options. Whereas the Matching Model may assist the user in selecting the appropriate techniques, the application of a workload technique may require the assistance of a human factors engineer. Consequently, there is an attempt to identify and locate expertise where ever possible.

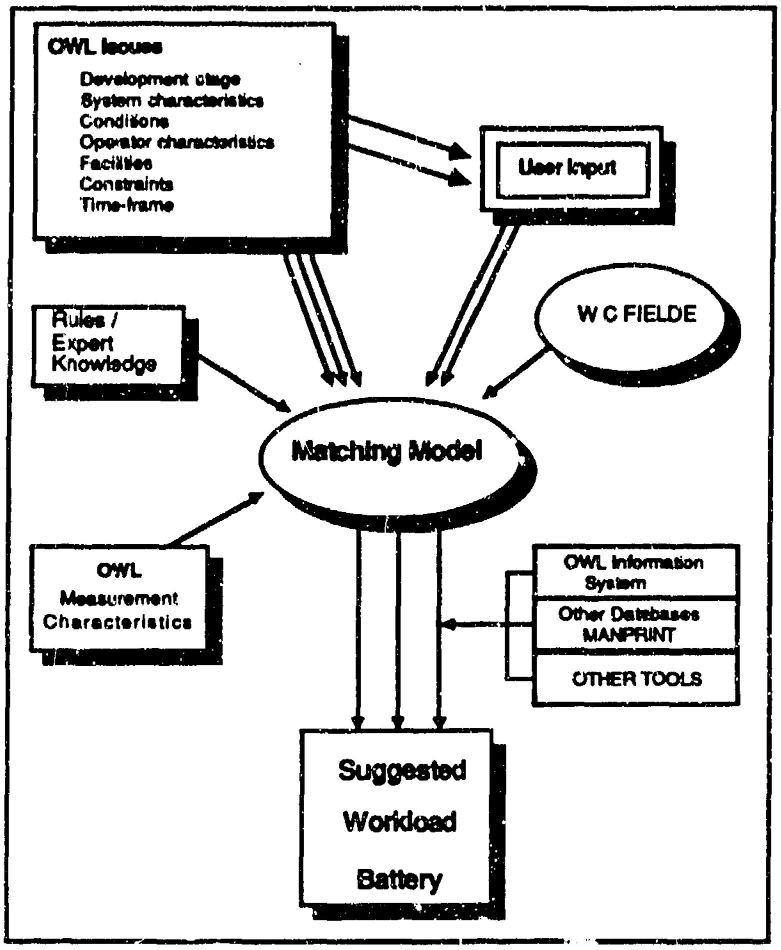


Figure 8-3. Illustration of the various components feeding into the Matching Model.

Matching Model Development

We began to construct our matching model by developing a list of relevant user questions, in a format appropriate for an expert system. The terminology follows that of an expert system shell and that of W C FIELDE. Each entry in the list consists of a:

- **Question** - to be answered by the user,
- **Reason** - basis for a decision rule (or set of rules) based on the answer to the question, and
- **Alternative User Responses** - possible user answers to the question.

Next our set of questions was compared with the issues covered in W C FIELDE. Differences were analyzed and appropriate revisions were made in our list. The result of this comparison process is the list of 23 questions presented below which form the basis of expert system development. Table 8-1 contains the set of operator workload techniques to be included in the Matching Model.

Question 1: What is the type of acquisition process the system is going through?

Reason: The selection of OWL techniques will depend on the type of acquisition process being used. A Non-Developmental Item may not have the concept phases and therefore both analytical and empirical techniques can be used from the beginning. Further, the time available for OWL analysis will vary.

Alternatives:

Traditional Materiel Acquisition Process (MAP)

ASAP

NDI

P3I

PIP

Question 2: If traditional MAP, what stage in the system acquisition cycle is the man-machine system currently in?

Reason: The first two alternatives are predominantly evaluated by analytical techniques. Usually, there is no hardware such as a simulator available to do any detailed testing with an operator in the loop. There are exceptions to this; for example, availability of generic simulators and rapid prototyping systems in which new displays can be installed and evaluated. The answer to this question has an impact on deciding which category of technique is more appropriate. While any technique can be used at any stage of development, typically, fewer possibilities exist during early stages of development. Flexibility in selection of various evaluative techniques increases as we go down the list. In some sense, the ease and cost of evaluation is also influenced, e.g., if only one simulator exists, scheduling time to perform tests will be more difficult than if several simulators exist. The capability of changing the workload through system design decreases as we go down the MAP list for cost reasons. Either of the first two alternatives will result in a suggestion of analytical techniques. Falling into the latter three categories does not eliminate the possibility of using analytical techniques.

Table 8-1. Complete list of techniques and measurement procedures for OWL

Alternatives

- **Analytical Procedures**
    - Comparison Analysis
        - Early Comparability Analysis
    - Expert operator opinion
        - Prospective rating scales – ProSWAT
        - Other
    - Mathematical models
        - Manual control models
        - Information theory model
        - Queueing theory
        - Other
    - Task analysis methods
        - Task Analysis *
        - HRTES
        - McCracken-Aldrich
        - Cognitive task analysis
    - Simulation models
        - Time line
        - Performance model (Card/Moran/Newell)
        - Micro Saint - network simulations
        - SWAS
        - SIMWAM
        - Human Operator Simulator (HOS)
- **Empirical Procedures**
    - Primary task
        - System Response
            - RMS Error
        - Performance related
            - Primary task speed *
            - Primary task accuracy *
        - Fine structure
        - Other
    - Subjective scales
        - Rating scales
            - Analytic Hierarchy Process
            - Cooper Harper *
            - Honeywell version of Cooper-Harper
            - Modified Cooper Harper *
            - Bedford *
            - SWAT *

* The measurement technique is included in W C FIELDE.

Table 8-1. Complete list of techniques and measurement procedures for OWL (continued).

Empirical Procedures (cont.)

- NASA TLX (NASA Bipolar) *
- WCI/TE
- Psychometric methods
  - Magnitude estimation
  - Equal interval
  - Paired Comparisons
- Specialized scales
  - Pilot Subjective Evaluation
  - Dynamic Workload Scale
- Questionnaires/Survey
- Interviews
- Other

Secondary task

- Embedded secondary tasks *
- Dual tasks
  - Sternberg Memory *
  - Mental math *
  - Shadowing *
  - Time estimation *
  - Communications *
  - Tracking *
  - Monitoring *
  - Choice RT *
  - Embedded secondary tasks *
- Other

Physiological & eye movements

- Heart rate *
  - HR variability (0.1 Hz) *
- Body fluid
- CFF
- Eye measurements
  - Eye point of regard - Eye movements *
  - Eye blinks *
  - Pupil diameter *
- EEG (brain activity)
- Evoked potential *
- Blood pressure
- GSR (skin)
- EMG (muscle)

Other techniques

- Video tape

* The measurement technique is included in W C FIELDE.

However, the emphasis may shift to empirical techniques and the data requirements are much more rigid due to the magnified cost of design changes. Here we need to focus on precise detailed problems. Part-task studies are quite useful to decide whether to make some hardware changes or possibly add decision aids.

Alternatives.

Pre-concept exploration

Concept exploration

Demonstration & validation

Full scale development

Production & deployment

Question 3: What is the time frame in which workload analysis must be complete?

Reason: Determine the impact of the analysis time frame on techniques selected, e.g., if time is short, then use subjective techniques for both analytical and empirical purposes.

Alternatives:

Less than a month

One to 2 months

Within 6 months

Within 1 year

More than a year

Question 4: What sort of system apparatus exists to assess workload during performance of primary tasks?

Reason: If no hardware exists, then we must rely on OWL analytical techniques, i.e., task analysis, simulation models, etc.

Alternatives:

Simulators

specific to current system

generic

Prototypes

Mock-ups

Production system

Question 5: What computer software facilities are available ?

Reason: If no software exists, then we must go to other techniques such as pencil and paper techniques, i.e., task analysis, but cannot use simulation models, etc.

Alternatives:

Computer simulation models

time line analysis

Micro-Saint

HOS

other

Data collection (interface software)

Data analysis

Statistical analysis packages

Question 6: What computers are available?

Reason: It requires a computer to run simulations. Different simulations run on specific machines and may not be compatible with other machines.

Alternatives:

Micro-computer (IBM-PC/AT) or compatible

VAX

Main frame

Other

Question 7: What sort of laboratory facilities are available for empirical work?

Reason: Some empirical techniques require specialized facilities or equipment. Primary and secondary techniques may require equipment to present tasks and record responses. Subjective techniques may use computers or paper and pencil. Physiological techniques may require equipment, such as sensors, to record physical responses.

Alternatives: Video, Audio, EEG, EKG, Pupil diameter measurement equipment, and Oculometer, etc.

Question 8: What staff support is available either in house or through another organization?

Reason: It is necessary to have the expertise (or the expert) available on the various topics.

Alternatives:

Expert operators on this or similar systems

Technicians, electronic, computer

Human Factors specialists

Personnel for testing in laboratory or field

Software developers - programmers

Statistical analysis support

Psychometric scaling and /or questionnaire expertise

Question 9: How much staff or manpower is available to do the OWL analysis?

Reason: Certain techniques (especially empirical) are very labor intensive. Certain techniques are more flexible than others in terms of manpower requirements.

Alternatives:

Less than 1 work week

Less than 1 work month

One to six work months

Six months to 1 work year

More than 1 work year

Question 10: Why is OWL assessment being done?

Reason: The reason OWL assessment is being done will influence the types of techniques used.

Alternatives:

MANPRINT requirement

Comparability analysis suggests chokepoint

Chokepoint already identified

Comparison of two (or more) candidate systems

Examination of individual differences

Question 11: What is the Mission Area (13 areas)?

Reason: Answers to this question will be helpful in directing the user to appropriate information already existing on workload evaluation. This breakdown will be helpful in tracing down comparable systems and may or may not be useful in the matching model. For instance, aviation systems have had considerable evaluation in the commercial arena by NASA, FAA, and by commercial aircraft companies. Other areas may or may not have a similar counterpart. This is also an attempt to use all information which may be available in other databases.

Alternatives:

Close Combat (heavy)

Close Combat (light)

Aviation

Air Defense

Combat Support, Engineering, & Mine Warfare

Combat Service Support

Fire Support and Target Acquisition

Nuclear, Biological, Chemical

Command & Control

Communications

Intelligence & Electronic

Special Operations

Combined Arms

Question 12: Is this a derivative system or a brand new one?

Reason: If it is a derivative system then the system can probably be tested in a generic simulator using the old system simulation model with mock-ups of the new operator controls and procedures.

Alternatives:

New

Derivative

Don't know

Question 13: What are the criteria against which to judge OWL with respect to overall man-machine system performance?

Reason: Need to know how the criteria were developed and to what they refer (this defines the boundaries of the criteria). Differentiate between system performance which includes the man and machine vs. human performance alone. A standard is needed to determine satisfactory system performance.

Alternatives:

Time requirements for mission objectives

Accuracy / Error requirements for mission objectives

Both time and accuracy.

Not identified

Question 14: What operating conditions (e.g., environmental conditions) and/or system usage factors need to be addressed or simulated by OWL assessment?

Reason: There are likely to be conditions under which the system cannot be tested even though they are conditions within which the operator would be under extreme stress, for example, battlefield conditions. These conditions will need to be addressed with analytical techniques even if the system exists.

(Part of the answer could be to highlight or alert the user to the existence of voids in the availability of techniques.)

Alternatives:

Deep Battle Environment
Covering Force Operations

Main Battle Area Environment

Rear Areas

Support Activities

Training

NSC Environment

Climatic Conditions - heat cold, etc.

Noise

Vibration

Question 15: Are individual operator differences important? That is, does the OWL analysis need to take into account the caliber and number of individuals available?

Reason: This question has to do with empirical evaluations. Test systems are often evaluated using top caliber operators. Even if individuals are successful, less capable operators may not be. This suggests getting ASVAB and other data available on the operators and comparing these test scores to the general level within the MOS.

Alternatives:

Yes

No

Question 16: What are the primary measures of human performance in the system?

Reason: This is an attempt to help the user define successful performance.

Alternatives:

Time requirements

Accuracy (or error) requirements

Both time and accuracy

Fine structure of behavior

Not identified

Question 17: What are the qualifications/characteristics expected for operators of the system? Do you need to consider manpower and personnel, and training issues?

Reason: This question has to do with MPT objectives. This is a step toward defining individual differences. If the analysis includes man-in-the-loop then we would recommend getting ASVAB, MOS test data, Skills Qualifying Test (SQT), and other available information. By knowing that, we would be able to infer what the dominant characteristics must be. (This question has to be stated in appropriate terminology, otherwise, the user may not know how to answer this very well.)

Alternatives:

Manpower requirements (e.g., crew size)

Personnel requirements - Aptitudes (e.g., coding speed)

Training - Skills and knowledge of soldier (e.g., time/accuracy requirements for performance of system tasks, knowledge of other systems interfacing with system to be developed)

Question 18: Has any OWL analytical analysis been done?

Reason: Analysis of workload is an iterative process, throughout the acquisition development cycle. It is important to determine whether system performance requirements were fulfilled and to identify the workload techniques used.

Alternatives: Analytical Procedures

Table 8-1 provides a list of these alternatives.

Question 19: Has any OWL empirical analysis been done?

Reason: When empirical analysis is possible (later in the development cycle) the information gained is very valuable to users and future OWL assessment. Empirical analyses, in general, have more face validity than analytical techniques because they are more grounded in reality.

Alternatives: Empirical Procedures

Table 8-1 provides a list of these alternatives.

Question 20: What operator performance characteristics are relevant to the particular man-machine system? (Universal operator behavior dimensions [Berliner, Angell, & Shearer, 1964])

Reason: We are interested in the categories of behaviors the operator must use. These questions can relate to the operators performance and to the ability of the system to meet performance characteristics, e.g., servicing targets, flying specific numbers of missions per day. (The user may not know how to answer this very well in the form given for the alternatives. It may be helpful to define the type of equipment and then to infer what the behavioral categories will be. If all of these alternatives are selected it will be necessary to break them out into subsets to deal with them more efficiently.)

Alternatives:

Perceptual

Mediational

Communication

Motor processes

(A complete list is provided in Table 8-2.)

Question 21: Can the operator be interrupted during a mission or are there blocks of time during the mission in which the operator can fill out forms?

Reason: Subjective measures require some time for filling out the rating forms. If the operator cannot be interrupted, then it is better to video tape the session and have the ratings completed later.

Alternatives:

Yes

No

It is possible to use video tape and get ratings later

Question 22: Does the operator have spare time to do other things at various points in the mission?

Reason: Secondary tasks may be used if there is some spare time.

Alternatives:

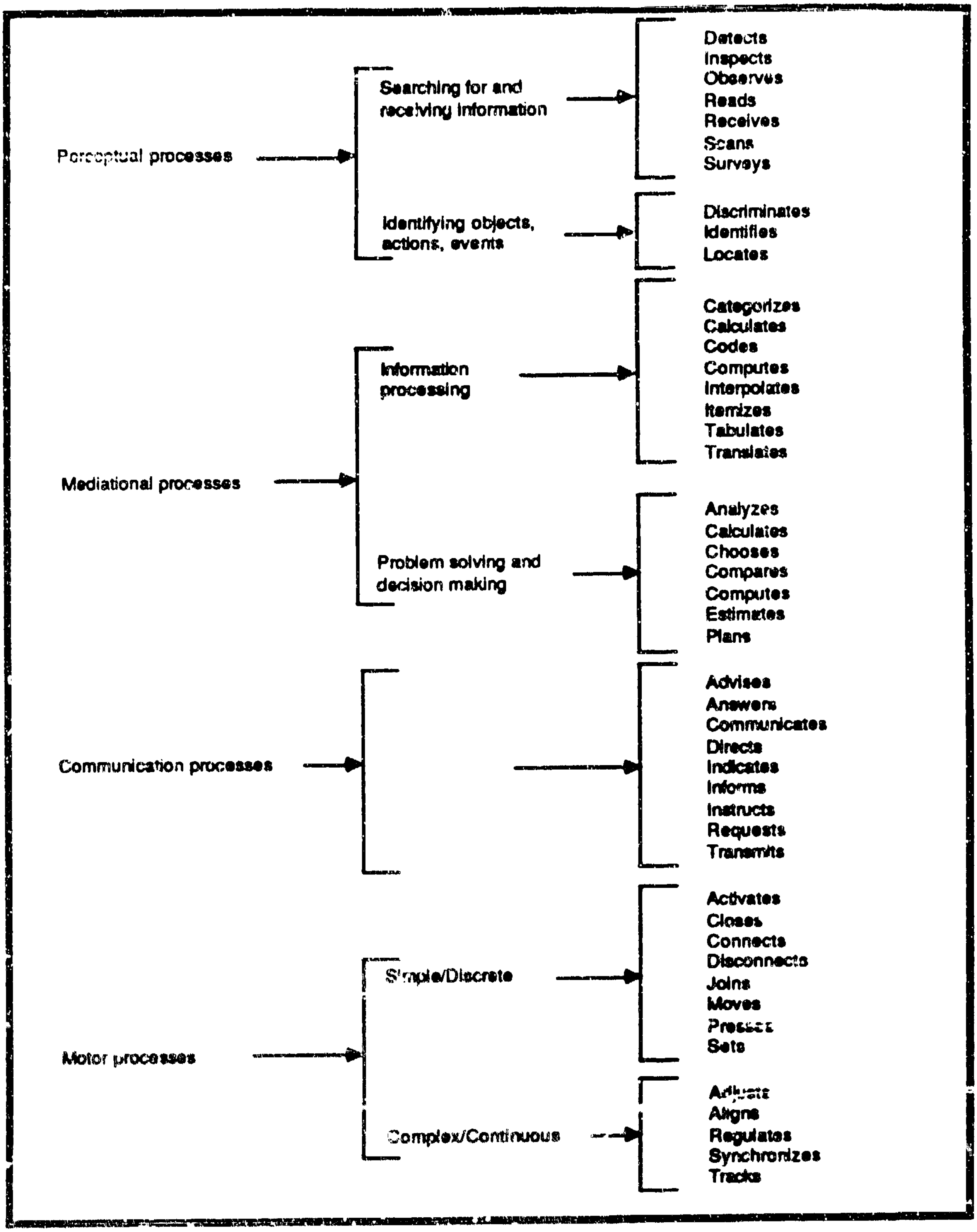
Yes

No

Question 23: What is the required duration of operator performance?

Reason: Again this is an important determinant of the types of data which can be collected. Short term and long term performance are different situations and require different treatment.

Table 8-2. Listing of Berliner et al. (1964) taxonomy of cognitive behaviors.



Perceptual processes	Searching for and receiving information	Detects Inspects Observes Reads Receives Scans Surveys
	Identifying objects, actions, events	Discriminates Identifies Locates
Mediational processes	Information processing	Categorizes Calculates Codes Computes Interpolates Itemizes Tabulates Translates
	Problem solving and decision making	Analyzes Calculates Chooses Compares Computes Estimates Plans
Communication processes		Advises Answers Communicates Directs Indicates Informs Instructs Requests Transmits
Motor processes	Simple/Discrete	Activates Closes Connects Disconnects Joins Moves Presses Sets
	Complex/Continuous	Adjusts Aligns Regulates Synchronizes Tracks

Alternatives:

Less than one minute

Less than an hour

One to two hours

Two to 8 hours

Sustained performance (over 8 hours)

Expert System Output

Possible Recommendations: The outcome and recommendations are selected from a comprehensive hierarchy of OWL techniques listed Table 8-1. Those which are addressed in W C FIELDE are noted with an asterisk.

Outcome Alternatives: The outcome possibilities are the entire set of techniques shown in Table 8-1.

Expert Systems

Two issues are considered in this section: What is an expert system and What are the reasons for an expert system?

What is an Expert System?

An expert system codifies the specialized problem solving expertise of an authority or, in some cases, many authorities to assist in solving complex problems in narrow domains. Expertise in a specific domain may generally be described as knowledge about the domain, the problems involving the domain, and the methods and approaches to solving the problems. The terms *expert system* and *knowledge-based* system are often used interchangeably to refer to artificial intelligence based systems that capture expertise in problem areas. In our approach, an expert system is considered to be a system consisting of two separate components.

- A knowledge-base representing the heuristics, facts, judgments, and experience about a selected problem domain.
- An inference processor which interprets the contents of the knowledge-base to infer conclusions toward a solution of the problem.

The separation of the knowledge from the inferential mechanism permits more flexible development and application, and more closely follows how humans deal with complex problem domains. Traditionally, expert systems are generated by a knowledge engineer who questions extensively an expert in a field to determine information and know-how about a selected topic, and translates the expert's knowledge into a knowledge-base. This knowledge-base construction is both the heart of and the main bottleneck to building an expert system.

The Reasons for an Expert System

There are a number of reasons for developing an expert system. Many of these reasons are listed in Table 8-3. While all of these reasons are relevant, the more important ones are: (a) communication of knowledge easily and efficiently, and (b) consistency and reliability.

Table 8-3. When expert systems pay for themselves (Van Horn, 1986).

- The expert is not always available, the expert is retiring, the expert is very expensive or rare
- A shortage of experts is holding back development and implementation
- Expertise is needed to augment the knowledge of junior personnel
- There are too many factors or possible solutions for a human to keep in mind at once, even when the problem is broken into smaller units
- Decisions must be made under pressure, and missing even a single factor could be disastrous
- A huge amount of data must be sifted through
- Factors are constantly changing, and it is hard for a person to keep on top of them all and find what is needed at just the right time.
- One type of expertise must be made available to people in a different field so they can make better decisions
- There is rapid turnover, a constant need to train new people. Training is costly and time consuming
- The problem requires a knowledge-based approach and cannot be handled by a conventional computational approach
- Consistency and reliability, not creativity, are paramount

Summary and Conclusions

This intent of this chapter is two-fold within the focus of OWL technique selection. First, the discussion lays out some examples for the immediate application of OWL techniques in the prediction and evaluation of workload. Second, this chapter outlines the general approach that needs to be taken for selecting OWL techniques. Twenty-three questions are presented which cover all major aspects of workload technique selection.

The general approach illustrates the seemingly complex set of considerations which must be addressed in selecting techniques. This general approach can best be implemented in a computerized expert system. Through this means, the development community has access to broad body of workload knowledge which is distributed and accessed in a systematic and efficient manner. Both the system developer and the workload analyst can identify easily the appropriate means to assess workload.

CHAPTER 9. CONCLUDING COMMENTS AND FURTHER DISCUSSION

The overall purpose of this report is to provide useful and practical information concerning operator workload (OWL). This information is used not only for the evaluation of existing Army systems but also for prediction of workload for future systems. Much of the material presented in the preceding chapters represents a fairly comprehensive review of how researchers and practitioners have defined and measured workload. In the review, we have presented and used traditional classification schemes (e.g., Hart, 1985a; O'Donnell & Eggemeier, 1986; Strasser, 1985) for organizing operator workload techniques.

A considerable amount of attention was devoted to explaining and defining workload. A number of definitions of workload as used by researchers were considered. Workload has been defined in terms of (a) the number of things to do, (b) the time required vs. the time available to do a task, and (c) the subjective experience of the operator. After considering a number of performance issues and these workload definitions, we developed the idea of a performance envelope in Chapter 2. The performance envelope is a generalized explanatory concept; it contains the foundations for each of the three definitions and allows for variations in performance within individuals and between individuals. Performance can be described as a momentary point in space within the performance envelope. We maintain that variations in operator workload, as well as other factors depicted in Figure 2-2, cause displacement of the operator within the performance envelope. The proximity of the individual's position to the boundaries of the envelope are indicative of the relative capacity to respond. It is this parameter of operator workload that we deem to be of major interest to system designers.

The main body of the report is a review and analysis of techniques that have been used for assessing OWL. These techniques were classified into two broad categories: The analytical category which contains predictive techniques that may be applied early in system design without the operator-in-the-loop and the empirical category which consists of operator workload assessments that are taken with an operator-in-the-loop, during simulator, prototype, or system evaluations. This analytical/empirical dichotomy is an important distinction in workload assessment. One objective of the present chapter is to provide some additional discussion on this distinction as well as some additional general comments related to operator workload.

The goal of workload assessment is to contribute to the processes that ensure acceptable system and human performance. It is useful to categorize OWL techniques as objective vs. subjective or primary vs. secondary, however, such categorizations tend to emphasize differences that are independent of the goal of workload assessment. In fact, such distinctions may serve to cloud the issue. We have often made the point in this report that multiple techniques should be used for a full OWL assessment. (Indeed, the dissociation of subjective workload assessment results from empirical performance is ample basis for this recommendation.) A more general and possibly more useful distinction is to consider operator workload from a cause and effect standpoint: this has some important implications for estimating workload. We want to draw out the implications by separating the determinants of operator workload from operator reactions to that workload. It will be argued that the cause and effect approach parallels the analytical/empirical distinction.

Operator workload research has tended to be atheoretical and this volume has been oriented unabashedly toward practical application of the techniques and concepts to Army systems. But there is nothing so practical as a workable theory. Accordingly, part of our discussion in this chapter revisits the workload model (Figure 2-2) and considers the review of workload techniques from a slightly different perspective from that generally presented.

The Determinants of Operator Workload

Factors both external to and internal to the operator will determine the extent of workload. The external factors include job requirements and job constraints that determine the workload of any job. The internal factors include the operator's own resources and capabilities. Together, these factors form the basis for analytical techniques; they are shown in Figure 9-1 and described below.

Job requirements. Job requirements will be a function of the types of tasks allocated to the operator and the rapidity of occurrence of various events to which the operator must respond. The types of tasks are important because they will determine the kinds of acts which an operator must be able to produce while doing the job. The rapidity with which events occur will determine the frequency and the time available to produce various acts. Further, the sequencing and timing of external events confronting the operator will vary from moment to moment. Because of this, the workload associated with a job will also change over time. The same observation could be made about the workload of a particular task or a particular act. Although it may be true that a single set of values could be estimated for the average workload of a job or task or act, it would be ignoring the fact that there are distributions of workloads for jobs, tasks, and acts.

Job constraints. Job constraints include the resources furnished to the operator (e.g., the design of the workstation, and the types of equipment and supplies the operator can use in performing the job). For example, the extent to which the workstation has been engineered for the human can impact significantly the workload of performing the job, and hence, the difficulties of various job-related acts. Because the working status of various equipment items and the availability of supplies and services may also vary from moment to moment, the job constraints will probably add to the overall variability of workload and performance distributions.

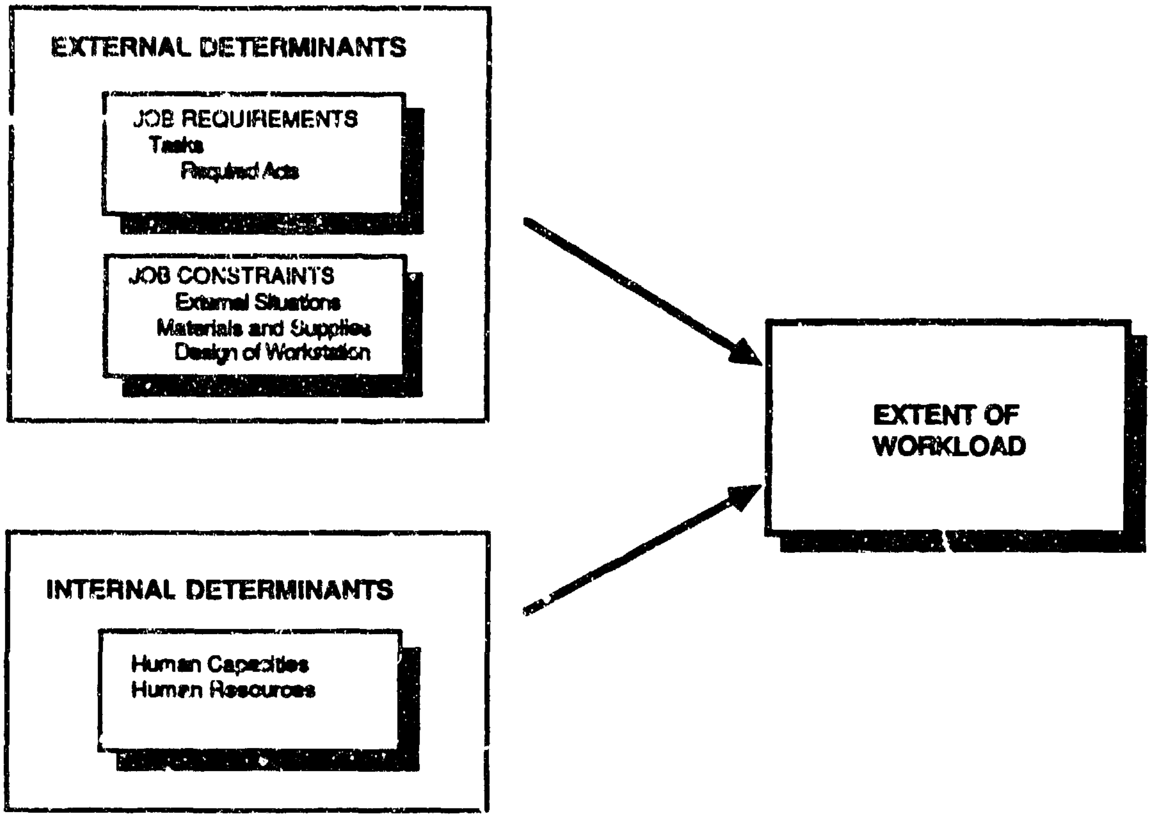


Figure 9-1. Illustration of the determinants of operator workload.

Internal Determinants. Factors within the operator determining the extent of workload relate to the various internal resources and capabilities that the operator carries into and applies to the job, tasks, and acts. There are, of course, tremendous individual differences in the resources and capabilities of potential operators. Appropriate selection, placement, and training of operators can be expected to result in greater suitability of internal capabilities and resources of operators assigned to a given job and also reduced variability of individual differences. Performance would be expected to improve as a result of a reduction in workload. However, it is unreasonable to expect that all individual differences can be eradicated by these means (e.g., Adams, 1987). Therefore, we must assume that even after appropriate personnel actions have been taken operators will vary in their capabilities to perform various jobs, tasks, and acts.

Interactions of External and Internal Factors. The extent of workload imposed on an operator will be a function of the external job requirements and constraints and the operator's internal resources and capabilities. These determinants differ from operator-to-operator, from day-to-day, and even from moment-to-moment. Perhaps more importantly, tasks can interact. Although it is perhaps a poor example because it could be memorized, the alphabetic/number interleaving task described in Chapter 2 is an example of detrimental task interaction Because of these variations, the estimation of extent of workload for a given job, or even a given mission or mission segment, is neither a simple nor a straightforward exercise. Certainly, workload cannot be evaluated by considering one task at a time without understanding the context of other current tasks that will be required.

Implications for Analytical Prediction of Operator Workload

Task Analysis. There are several implications of the cause/effect distinction for the prediction of workload. For example, enumerating the various operator tasks can probably be accomplished with relative ease. Determining the frequency, sequencing, and especially the interactions of those tasks is far more difficult. It requires careful and accurate determination of the external events which will probably occur for a variety of mission situations as well as the careful and accurate determination of the probable external resources that will be available to the operator.

Obtaining accurate estimates of the expected frequencies and sequences of the operator's tasks is, however, only the starting point for workload analysis. One of the goals of traditional task analysis was to arrive at estimates of the kinds of acts (e.g., perceptual, cognitive, psychomotor, communications) required by those tasks. A second goal was to arrive at estimates of the times required by those acts (and hence, by the tasks in which those acts occur). The times to accomplish various perceptual and psychomotor acts will ultimately depend on the operator's workstation; accurate estimates of some times are not possible without first determining the layout of the workstation. There remains much controversy over these time estimates (e.g., Holley & Parks, 1987), especially when they are obtained in a subjective manner and the estimators are not required to specify what assumptions they have made. To ameliorate this problem, most analytical techniques have recommended using SMEs (e.g., actual operators) in a standardized, structured estimation process.

Performance Models. For obvious reasons, SMEs are generally preferred over novices. Some of the techniques reviewed in Chapter 3 (e.g. HOS) reduce the use of SMEs to (a) describing the detailed steps in each task and (b) the likely design of the crewstation displays and controls. HOS assumes that SMEs can be relied upon to arrive at sufficiently valid descriptions of this type of information, but it does not make the additional assumption that SMEs are necessarily good evaluators of either the acts or the times that will be required by the detailed steps in each task. Indeed, we know that operators do not do well in these regards (Spady, 1978a). Instead, HOS relies on a fairly complex, computerized human operator model to accomplish those latter functions.

Determining accurate estimates of which operator acts will be required and when they will be required are, themselves, complex problems. But, in reality, they are merely the beginning steps for predicting human performance and estimating operator workload. The time needed to perform the acts and the accuracy with which those acts can be accomplished will be a function of the capabilities and resources of each individual operator. Thus, one can expect a distribution of relative workloads for different operators just as one can expect a distribution of relative workloads for different missions and scenarios.

The time and accuracy to perform any given act will be dynamically changing from moment to moment. Because the ultimate measure needed for workload is related to how close the individual is coming to the edge of his acceptable performance envelope, estimating workload remains a highly challenging problem.

Measuring Operators' Reactions to Workload

The preceding section contained discussions about the various determinants and causes of workload, on the analytical side of workload analysis. In that approach, one examines the conditions under which an operator will be required to perform tasks and on that basis, arrives at estimates of what the performance and workload should be. On the empirical side, we examine various types of operator reactions to performing a job and, based on those reactions, arrive at estimates of what the workload must have been. In this section, therefore, we will discuss the various outcomes and effects resulting from workload. However, rather than use the more traditional four-level taxonomy presented earlier, it is suggested that the measurement of outcomes of workload can be described more parsimoniously in tripartite terms of the operators' (a) job-related acts (primary and secondary task measures), (b) concomitant physiological changes, and (c) subjective reactions engendered as the operator attempts to perform the assigned job. Figure 9-2 illustrates these effects of operator workload.

Job-related Acts. Job-related acts are synonymous with job requirements and required acts described in the discussion of workload determinants. Estimating when certain acts occur can be done using either subjective or objective techniques. Some of these acts are observable and can be measured in highly objective fashion. For example, head and eye movements and visual fixation can be determined objectively by measurement of the eye. Limb movements, grasping and manipulating control devices, and use of speech for messages can be also measured and timed very accurately by the primary and secondary task measurement techniques. However, not all human job-related acts are directly observable. EEG data indicate that internal events are continually occurring within the brain, but they provide little information on what specific events are occurring. To a large extent ECPs and the variability in heart rate may be considered as preliminary attempts to measure indirectly the occurrence of those various internalized activities which are not directly observable.

Physiological Changes. Physiological changes represent the second type of reaction to the workload an operator confronts. There are two broad subtypes of physiological changes: Momentary and long-term. The momentary changes are represented by ECP techniques, pupil responses, eye

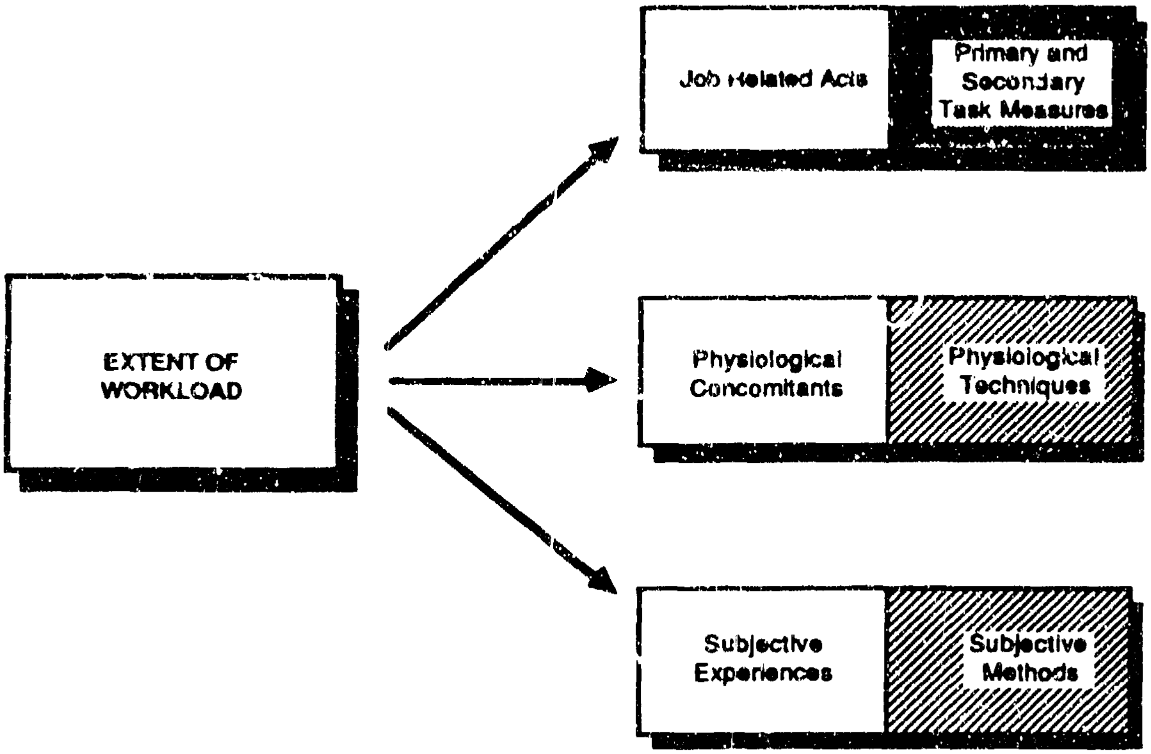


Figure 9 2 The measurable reactions to workload. The assessment techniques are shown in the shaded areas.

movements, and the like. These momentary changes have been well documented. There also is little question in the long-term that a variety of biochemical byproducts are generated as the operator goes about performing a job. The increase or decrease of certain chemicals in the body may be related to the depletion and recovery of some of those resources needed to perform various acts. What we are speaking of in the long-term case are not the various physiological indicators that a task-related act has occurred, but rather the concomitant physiological changes occurring because internal resources are being depleted. Several of these types of physiological changes are discussed in Chapter 7. Because of the complexity of the varying time delays involved in the underlying physiological processes, the detectable changes have little diagnostic value at present for determining precisely which acts were related to those changes

Operator Experiences. Subjective experiences are the third type of reaction to the workload confronting an operator. We know that some types of experiences (e.g., anxiety, fear, fatigue, confusion, frustration, anger, failure) encountered in work situations are correlated with various concomitant physiological changes. We also know that some types of experiences may be correlated with the operator's level of performance. Thus, there is an obvious overlap between an operator's reactions to workload that we classify as an experience and the other two workload effect categories. However, the overlap is not perfect and sometimes the correlation is even negative. This has been called dissociation by various investigators (e.g., Derrick, 1988; Yeh & Wickens, 1984).

A related problem is that the kinds of experience frequently reported by workload researchers are correlated among themselves. Thus, it is likely that a person who reports experiencing confusion may also report feeling frustrated as well. If experiences are important effects of workload, we need to isolate their independent dimensions. Subjective experiences are also likely to undergo continuous changes during a mission. If one waits until the end of a mission to collect data on subjective experiences, it is possible that earlier experiences may have been forgotten. It is also likely that subjective reports will be influenced by the perceived outcomes of the mission. For example, when a mission is ultimately successful, operators may tend to play down the importance of an earlier feeling of confusion. By the same token, if the outcome of a situation is a failure, the operator, even though he never actually felt that way during the mission, may now recognize that he must have been confused. This could lead to the operator reporting confusion even though he never really felt that way.

A final point is that information is sometimes solicited under the guise of the operator's subjective experiences when, in reality, the information we want is the operator's evaluation and judgment concerning the goodness of the design of the system. A relevant question, then, is whether operators would arrive at the same conclusions about a system design without ever asking about their experiences. From a designer's point of view, it will almost always be more informative to know specifics about tasks or components of the system with which operators experience difficulty, than just to know that he experienced an overload. Additional useful and diagnostic information can be collected by eliciting direct information from operators about how the system might be improved.

The Relations between the Determinents and Effects of Workload

It should be possible to estimate the extent of workload by examining either the determinants or the effects of the workload. For a given situation and individual, one would expect the two approaches to yield similar answers to the question of what the extent of workload was. Indeed, it is entirely possible that they will not. The major reason can be described in terms of the parallel distinction between analytical and empirical techniques and the capability of predicting vs. measuring performance. The most obvious overlap between the two approaches lies in the area of job-related tasks and their performance by a given operator. That is, a full understanding of the task demands, situational constraints, and capabilities and limitations of a given operator should yield acceptable predictions of what acts will occur, when they will occur, and how well they will be executed in that situation. This desired agreement between predicted and actual acts is one of the major goals of the development of human performance prediction models. It may be argued that validation of all of the predicted acts cannot be obtained because some acts are simply not observable. There is, however, a partial answer to this objection. If there is acceptable correspondence among the predicted and actual observed acts and events, then the model is probably accounting for the times required by the unobservable acts.

The determinants approach to assessing OWL suggests that various internal resources must be being depleted as acts take place. The effects approach suggests that some of the detectable physiological changes might be indicative of the fatigue and recovery of those same internal resources. Thus, there is a second way in which the two approaches might be shown to correspond. It might well be, however, that some internal physiological changes cannot be detected or observed with the present technology. Theoretically, the depletion and recovery of various internal resources are responsible for changes in the level of performance of the job-related acts. Thus, if the human performance model correctly predicts when the performance of various acts will degrade or improve, then it can be assumed that the depletion and recovery of internal resources is being accounted for. Current human performance models do not include provisions for the depletion and recovery of internal resources needed for the production of various acts. However, there is nothing that prevents that concept from being included.

The determinants approach, unlike the effects approach, fails to consider an operator's internal experiences. If some of the subjective experiences are considered to be the results of the operator's perceptions, then they also could be modeled. For example, anxiety and fear could be assumed to occur when the simulated human assesses the situation in which he finds himself as being threatening. Confusion could be assumed to occur when the simulated operator cannot solve problems as rapidly as he normally can or when processed information (perceived and/or recalled) is found to be contradictory. Feelings of physical or mental fatigue could be assumed to occur when the corresponding acts have been required over a sustained period of time. Even feelings of being overloaded could be assumed to occur whenever the model of human performance indicates a recognition of job demands outpacing available time. Thus, it is conceptually feasible (though a major undertaking) to construct human performance models that would also include the generation of subjective experiences as well as the performance of the assigned tasks. As it would have to include various introspective tasks along with the job-related ones, such a model would be more complex that existing ones. It might be, however, that ultimately such models would be more accurate in predicting actual job performance because they could account for the internal motivations of the operator.

The Need for Further Validation

The cause (determinant) and effect (reaction) approach is nothing more than looking at two sides of the same coin. Researchers have done this implicitly in developing and applying analytical techniques by often using data from empirical techniques. But they have simply not gone far enough. Too few of the analytical techniques have been sufficiently validated. Without thorough validation, we do not know if we have a good, practical technique or an interesting, untested theory. The lack of validated analytical information on OWL suggests looking at both sides of the coin. As a result of converging operations (Chapter 2) a clearer picture can be obtained from several uncertain views. This is the predominant rationale for our advocacy of OWL technique batteries.

Future Directions

Having considered virtually every workload assessment technique and thereby having obtained a rare overall perspective, we would be remiss not to highlight several gaps in the technologies and content of workload research. Of course, many things need to be done to create more applicable and validated workload tools and techniques. Analytical techniques in particular represent an area containing many technological gaps. This has been considered in the discussion of Chapter 3 as well as in this chapter. Two major areas of OWL research that need further study are: multiple task performance and individual differences. Each of these topics not only has impact on operator workload evaluation, but also on performance and other areas of MANPRINT concerns.

Estimating Workload for Multiple Tasks and Multiple Situations

A good deal of the laboratory workload research has dealt with single or dual task experiments occurring within a single or possibly two different sets of task conditions or situations. Although the results of that research have been interesting, most Army jobs of interest have multiple ongoing tasks. Further, as pointed out earlier, even though the nature of an operator's tasks may be similar from one day to the next, the relative difficulty of the various mission situations confronted may change significantly on a day-to-day basis. Figure 9-3 illustrates the two dimensions of number of concurrent tasks and number of different situations in which the tasks occur. This figure, of course, is an abstraction. In reality separating discrete elements of tasks or situations may be difficult.

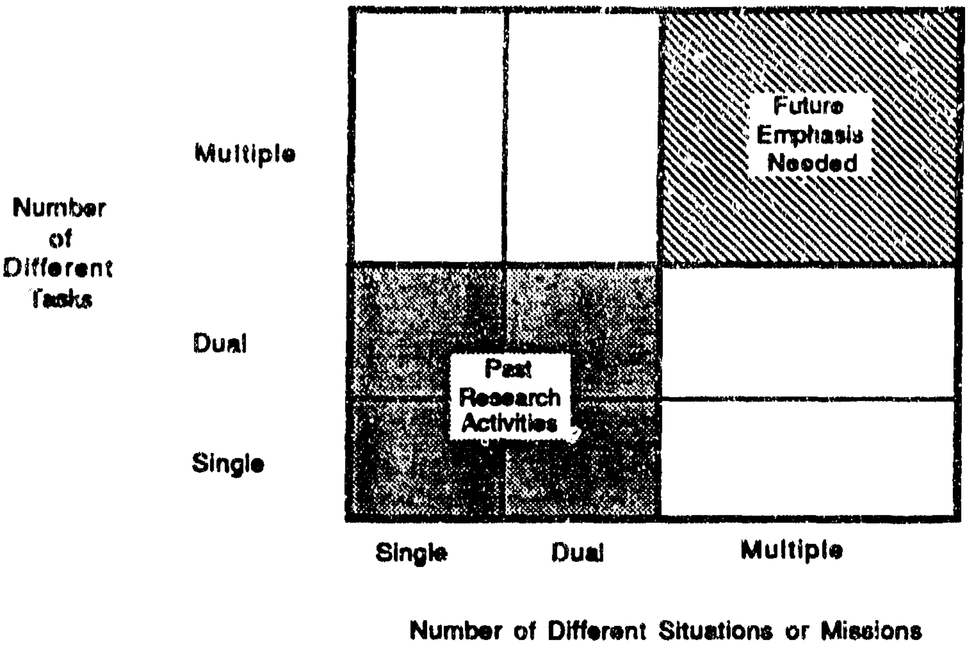


Figure 9-3. Relation of number of tasks and number of situations.

Despite the fact that most empirical data collected on operator workload come from the lower left cells of Figure 9-3, the major interest for future workload research will be the upper right multitask, multisituation cell. In many experiments, subjects have been required to perform only a single task. In this case, it is relatively easy to determine how well the subject has performed the job. When a second task is added to the job, as in the case of dual-task studies for example, it is much less clear how one should evaluate the overall performance of the job. Yet, it is clear that job performance and operator workload cannot be evaluated by examining only the performance on the primary task ignoring time-sharing requirements. Rather, overall performance on all ongoing tasks must be considered in arriving at estimates of workload. How this is to be done is one of the most challenging issues, not only for workload assessment, but also for overall performance prediction and evaluation.

An advantage to collecting data on multiple situations is that, not only can measures of interesting parameters for each task be obtained for each situation, but they can be compared across the various situations. For example, our discussions in earlier chapters indicated that changes in the frequency or extent of certain kinds of acts from one period to another might be indicative of the operator applying a different set of performance rules for the same acts in different situations. The technique of adding a secondary task can best be understood in the context of changing the situation in order to see how it affects performance on all of the other tasks. The issue of predicting the impact of additional tasks on overall job performance is, of course, central to the whole problem of allocating tasks to an operator. The concept of manipulating the difficulty of any task to see how it will impact overall job performance is also a useful technique. Adding new tasks or increasing the difficulty of existing tasks are alternative techniques for determining the location of the operator in his performance envelope. In the process of evaluating operator workload, we do not yet fully understand where the boundaries of acceptable workload are for the human. Incrementally adding to the workload until performance begins to deteriorate or to breakdown is similar to methods used in the physical sciences for testing the tensile strength of various materials.

We have stressed the importance of not only estimating what operator performance will be under a variety of expected mission scenarios, but also knowing how close it will come to the boundaries of unacceptable performance during those missions – even if every operator's performance were completely acceptable. The purpose in doing this is to understand better the margin of error in a proposed design of a new system. That margin of design error is especially important because future demands of any job may well be more difficult than originally anticipated and new tasks may have to be added to counteract technological advances or doctrinal and tactical changes in the employment of hostile forces.

The application of improved knowledge about performance in multitask situations will clearly benefit the system designer and impact not only workload evaluation but also a variety of MANPRINT issues. The designer will benefit by being able to improve designs and optimize task allocation. The trainer will benefit by having a better understanding of performance rules and which components need more emphasis. The trainer will also benefit by being able to teach time-sharing skills. The suggested approach is clearly interdisciplinary. The need for considering diverse sets of data from neuropsychology, from individual difference research, from performance research, from human modeling and artificial intelligence, and from mathematical modeling is simply too much for any single researcher to master.

Attention and Switching Among Tasks. A general conclusion from this review is that a full understanding of operator workload will begin to emerge only when sufficient workload investigations have emphasized multiple tasks and multiple situations. The suggestion of looking at multiple tasks and multiple situations is a general plea, however, and we can be more specific. Because multitask situations are common occurrences for an operator currently and may well become even more common, we need more information and data about multitask performance and an understanding of the relations among and impact of individual tasks on multitask performance. In particular, the issue of time-sharing abilities comes into focus in this context.

The importance of the interactions of two or more tasks on performance cannot be overestimated. Mental workload often increases when two or more tasks are to be performed concurrently. This is certainly not surprising from several different theoretical approaches. We prefer an explanation involving attention switching for the following reasons. First, if one assumes that operators can consciously attend to only one thing at a time, the multitask situations require operators to decide which task should be to only one thing at a time, the multitask situations require operators to decide which task should be attended to at various points in time. This additional mental task clearly does not exist in single task situations. Such decisions, especially in rapidly unfolding combat situations, are far from trivial. The operator may well feel torn between working on several critical tasks, each of which is currently demanding his attention. Second, especially when tasks are considered to be approximately of equal importance, multitask situations may result in frequent interruptions of the current task to determine the need to work on the others. Even if the task is considered the most critical, the time and effort to evaluate the status of the other tasks takes mental time and effort. Third, the interruption of a task means that some time elapses during which the interrupted task is not consciously attended to. When the operator returns to the interrupted task, he may be surprised at what he now finds. The actions required may be somewhat different from what he had anticipated thus requiring yet additional mental effort. Finally, the continual switching among several tasks may require some additional time to reestablish the contents of the operator's working memory with the current situational data and the procedural rules for the task currently being worked on.

There are a number of experiments supporting the attentional switching conceptualization. Mewhort, Thio, and Birkmayer (1971) used dichotic listening and showed, independently of other factors, that the number of required switches had a dramatic impact on recall. In a different context, Weichselgartner and Sperling (1987) concluded that attention consists of two partially concurrent processes. One is a fast, effortless, automatic process (on the order of 100 ms) and the other is slower, effortful, controlled process (on the order of 300-400 ms). The faster process is affected by manipulations often considered as parallel processes that are independent of task difficulty while the slower is affected by variables typically considered as serial processes and related to task difficulty as well as training and practice.

Although data bearing on attention and switching problems are available from dual task investigations, many of the experiments reported in the literature have used tasks (both primary and secondary) that have little similarity with real world tasks. Attentional decisions as to which of the two tasks to work on in those situations are trivial when compared to most real-world situations of interest. Indeed, conclusions drawn from those types of investigations may simply be irrelevant to the real problems confronted in designing complex human-machine systems. Our earlier statement that future experiments should investigate multiple tasks performed in multiple situations or under multiple conditions includes the collection of relevant data for understanding attention switching problems.

The Need for New Metrics. As more realistic multitask multisituations are investigated, the issues of performance and workload trade-off and how they can be handled, will become more and more apparent and more pressing. New metrics are needed to facilitate more precise predictions about the trade-offs. One metric proposed is the performance operating characteristic (POC) (Norman & Bobrow, 1975). The POC is a way of representing the data obtained from two tasks done individually and in combination. Under the multiple resource theory (Navon & Gopher, 1979), Wickens, Mountford, and Schreiner (1981) have developed a normalization scheme which they claim provides such a metric. Essentially, their recommendation is to normalize dual task performance to single task performance. Kantowitz and Weldon (1985), however, have shown some of the dangers of using such a procedure. Through simulation, they have shown that erroneous conclusions could be drawn from the application of such a transformation. Further clarification has been offered by Wickens and Yeh (1985). Until these issues are settled, the POC may be a useful way to present data but its application should be extended only with care.

It may turn out, as suggested by Pachella (1974), that speed/accuracy trade-offs and other similar performance trade-offs should be handled with weaker scales of measurement (e.g., ordinal) and not the interval scales currently attempted. Other mathematical techniques may also be useful such as correlation, conjoint measurement, and factor analysis. Regardless of the ultimate nature of the metrics needed, it is obvious that much work currently remains to resolve this problem.

Individual Differences, Performance, and Workload

Our account started in Chapter 1 with a quote from a little book that dealt with testing for individual differences and it is fitting to close with some comments on the same topic. The usefulness of personnel testing during World War I was clearly demonstrated and documented in that 1920 reference; such testing has continued to the present. Although such information is highly useful, it is useful only in a broad sense. It does not provide the detailed information needed to predict operator performance precisely. We can ill afford to build a system *and then* determine whether soldiers can operate it. To use analytical techniques in a more beneficial manner to determine performance before the system is built, it is necessary to have a considerable amount of detailed information. In many ways, this approach compliments the multitask approach.

It is not that individual differences have been ignored in the experimental literature, they have been consciously set aside in favor of examining population means. (This is by no means a new point [e.g., Noble, 1961; Ozier, 1980].) About the only information available from many experimental reports related to individual differences is the information contained in the error terms and the subject term of analysis of variance or in the means and standard deviations which are sometimes presented. Further, much of the research conducted in university laboratories has dealt with a highly restricted population. Accordingly, even if information was presented about individual differences, one usually does not have information about differences for the population of Army operators. While most investigators support the importance of individual differences, one can find only a few volumes (e.g., Eysenck, 1977; Miles, 1936) that deal with individual variations in an experimental context and provide data needed for the analytical techniques. Fortunately, the situation is not as stark in other areas of research.

Any consideration of individual differences and mental workload lands one squarely in the domain of intelligence. Many years ago, the neuropsychologist Karl Lashley (1929) outlined several major theoretical positions on intelligence. Two of these general theories are extant: the general plus special abilities theory (e.g., Spearman, 1927) and the specific abilities theory (e.g., Thurstone, 1938). The general theory holds that there exists one general intelligence factor plus some ability specific to the test used. By contrast, the specific theory holds that intelligence is the algebraic sum of a number of diverse capacities. There are, of course, many variations of these two classes of theory; the theory subscribed to can have tremendous implications for the approach taken toward individual difference research. Much of the research in individual differences in abilities has utilized factor analytical approaches.

Emphasis on Underlying Acts. Akin to traditional factor analysis are several theories and approaches which emphasize the various capacities and resources that underlie performance on many different tasks. Work on attention has emerged from information processing and cognitive theories about behavior Coupled with the attention work is the evaluation and identification of mental acts involved in performance. Typically these approaches have not focused on individual differences but there is no reason why they cannot be extended. Navon and Gopher (1979) postulated that different amounts and types of resources are required for different task combinations (cf. Navon, 1984). Wickens (1984), building on the information processing approach, has formulated this idea into a relatively few general dimensions (e.g., verbal, spatial, auditory, visual, speech, motor) to deal with the multiple task problem. There are a number of other ways of examining behavioral detail associated with mental acts. Researchers have generated a considerable amount of data relevant to the issue using a number of approaches, including perceptual processing (Garner, 1974), brain damage (Luria, 1966), skill learning (Adams, 1987), and the nature of intelligence (Guilford, 1967).

In his Underlying Internal Processes (UIPs) theory, Wherry, Jr. (1986) emphasizes individual differences and postulates that most differences in task performance are attributable to the number of times different UIPs must be invoked for a given task and how fast and how accurately different UIPs are performed by those individuals. Based on the established moment to moment reliabilities of task performance, he maintains that the time and accuracy of given UIPs within individuals must also remain fairly constant. He presents a methodology by which the number and nature of the different UIPs required for given tasks can be identified by the analysis of the correlations among task completion times across many variations of the task of interest. His analysis also permits the estimation of the individuals' capabilities for the identified UIPs.

These approaches have much in common. They share with traditional factor analysis approaches to intelligence research the concept that individual differences in task performance are attributable and explainable by understanding the differences in humans' capabilities to perform various kinds of underlying mental acts. As such, they also share much in common with the required acts as explained in our recasting of analytical and empirical workload estimation approaches. Thus, we also conclude that much more attention must be paid to quantifying individual differences in underlying capacities if workload estimation is to progress to a mature and useful technology.

Skills and Performance Rules. In addition to the approaches already mentioned, there are several other directions individual difference research might take. One of these is in the acquisition of skill and how tasks and acts become automated. The other is in the performance rule/strategy domain. Individuals differ not only in their underlying capabilities, but also in the knowledge they may bring to bear on various problems. Ozier (1980) has shown clearly the role of performance rules in free recall. The differences found are quite striking even with a restricted population of college students. Ozier suggests that these organizational rules (or strategies) are independent of scores on several intelligence tests. Whether application of performance rules is independent of or related to general abilities is of tremendous practical importance to operator workload issues. Independence implies no predictability and without predictability, all of our performance models are inadequate.

Project A. Yet another approach to individual differences is represented in Project A. This is a large scale program undertaken by the Army Research Institute to supplement the ASVAB for the purpose of improving the prediction of successful performance in both training and on the job. Peterson (1985) provides an overview of the steps taken to develop additional personnel tests. The overall program involves not only developing new tests but also validating these tests against criterion soldier performance. When completed, the database will contain a considerable amount of information of relevance to performance. Of particular importance is the fact that some of the new tests being developed are computer-based performance tests in which information being evaluated by subjects can be dynamically changing and performance patterns and performance times can be measured. Thus, it may be possible to assess underlying processes not testable by the typical paper and pencil methods.

And Finally, it really is an Elephant!

It is difficult, perhaps impossible, to summarize a volume of this size in a few well chosen sentences. The reviews of workload definitions, techniques, and approaches represent a massive effort rarely undertaken. After having the opportunity to examine the great diversity of those definitions, approaches, and workload estimation techniques, we have been struck by the fact that we have, indeed, been looking at the same elephant.

One objective of this final chapter was to provide a further synthesis of the materials from the preceding chapters. We have attempted to illustrate and clarify the very real overlap between traditional analytical and empirical workload techniques with our discussion of the causes and effects of workload.

A second objective was to assess briefly past research efforts and indicate future directions that will strengthen the body of knowledge upon which more coherent and encompassing models of operator workload can be constructed. To this end, we have advocated a much greater emphasis on multitask and multisituation investigations as well as greatly expanded interest in individual differences.

A final objective of this chapter was to emphasize that determining the extent of workload is not an end in itself. It is, however, a necessary step in determining the position of the operator within their own performance envelopes which in conjunction with the nearness of the boundaries of those envelopes, provides an indication of the operator's momentary relative capacity to respond. It is this parameter, more than any other, that system designers require if they are to build adequate man-machine systems.

The importance of understanding the level of operator workload is clear: High workload may result in unexpected and undesirable performance changes. The operator may shed tasks, be unable to perform them, or in some other way fail to perform acceptably. In one form or another, rightly or wrongly, the operator will adapt. Without such consideration, the incorporation of MANPRINT concerns into the design of systems will continue to be problematical.

REFERENCES

Aasman, J., Mulder, G., & Mulder, L. J. M. (1987). Operator effort and measurement of heart-rate variablity. *Human Factors, 29*, 161-170.

Acton, W. H., & Colle, H. (1984). The effect of task type and stimulus pacing rate on subjective mental workload ratings. In *Proceedings of the IEEE 1984 National Aerospace and Electronics Conference* (pp. 818-823). Dayton, OH: IEEE.

Acton, W. H., & Crabtree, M. S. (1985). Workload assessment techniques in system redesign. In *Proceedings of the IEEE 1985 National Aerospace and Electronics Conference* . Dayton, OH: IEEE.

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin, 101*, 41-74.

Albery, W., Repperger, D., Reid, G., Goodyear, C., Ramirez, L., & Roe, M. (1987). Effect of noise on a dual task: Subjective and objective workload correlates. In *Proceedings of the National Aerospace and Electronics Conference*. Dayton OH: IEEE.

Aldrich, T. B., & Szabo, S. M. (1986). A methodology for predicting crew workload in new weapon systems. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 633-637). Santa Monica, CA: Human Factors Society.

Aldrich, T.B., Craddock, W., & McCracken, J.H. (1984). *A computer analysis to predict crew workload during LHX scout-attack missions: Volume 1.* Unpublished technical report. Ft. Rucker, AL: U.S. Army Research Institute.

Allen, M., & Yen, W. (1979). *Introduction to Measurement Theory.* Monterey, CA: Brooks/Cole Publishing Company.

Allen, R. W., Jex, H. R., McRuer, D. T., & DiMarco, R. J. (1975). Alcohol effects on driving behavior and performance in a car simulator. *IEEE Transactions on Systems Man and Cybernetics, SMC-5*, 498-505.

Allport, D. A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology, 24*, 225 -235.

Alluisi, E. A., & Morgan, B. B. (1971). Effects on sustained performance of time-sharing a three-phase code transformation task (3P-Cotran). *Perceptual and Motor Skills, 33*, 639-651.

Angus, R. C., & Helsgrave, R. J. (1983). The effects of sleep loss and sustained mental work: Implications for command and control performance. In *Sustained intensive air operations: Physiological and performance aspects.* AGARD Conference Proceedings No. 338.

Armstrong Aerospace Medical Research Laboratory (1987). *Subjective Workload Assessment Technique (SWAT): A User's Guide.* Dayton, OH: AAMRL, Wright Patterson AFB.

Attneave, F. (1959). *Applications of information theory to psychology.* New York: Holt.

Baddeley, A. D. (1966). The capacity for generating information by randomization. *Quarterly Journal of Experimental Psychology, 18*, 119-130.

Bahrick, H. P., Noble, M., & Fitts, P. M. (1954). Extra-task performance as a measure of learning a primary task. *Journal of Experimental Psychology, 4*, 299-302.

Bainbridge, L. (1974). Problems in the assessment of mental load. *Le Travail Humain, 37*, 279-302.

Bainbridge, L. (1978). Forgotten alternatives in skill and work-load. *Ergonomics, 21*, 169-185.

Baron, S. (1979). A brief overview of the theory and application of the optimal control model of the human operator. In M. C. Waller (Ed.), *Models of human operators in vision dependent tasks*. NASA Conference Publication 2103. Washington, DC: NASA.

Baron, S., & Levison, W. H. (1977). Display analysis using the Optimal Control Model of the human operator. *Human Factors, 19*, 437-457.

Baron, S., Zacharias, G. Muralidharan, R., & Lancraft, R. (1980). PROCRU: A model for analyzing flight crew procedures in approach to landing. *Proceedings of the Eighth IFAC World Congress*. Tokyo.

Bateman, R. P., & Thompson, M. W. (1986). *Correlation of predicted workload with actual workload measured using the subjective workload assessment technique*. SAE AeroTech.

Bauer, L. O., Goldstein, R., & Stern, J. A. (1987). Effects of information processing demands on physiological response patterns. *Human Factors, 29*, 213-234.

Beare, A., & Dorris, R. (1984). The effects of supervisor experience and the presence of a shift technical advisor on the performance of two-man crews in a nuclear power plant simulator. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 242-246). Santa Monica, CA: Human Factors Society.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276-292.

Becker, C. A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 556-566.

Bell, P. A. (1978). Effects of noise and heat stress on primary and subsidiary task performance. *Human Factors, 20*, 749-752.

Benson, A. J., Huddleston, J. H. F., & Rolfe, J. M. (1965). A psychophysiological study of compensatory tracking on a digital display. *Human Factors, 7*, 457-472.

Bergeron, H. P. (1968). Pilot response in combined control tasks. *Human Factors, 10*, 277-282.

Berliner, C., Angell, D., & Shearer, D. J. (1964). Behaviors, measures, and instruments for performance evaluation in simulated environments. Paper presented at the *Symposium and Workshop on the Quantification of Human Performance*, Albuquerque, N.M.

Biferno, M. A. (1985). *Mental workload measurement: Event-related potentials and ratings of workload and fatigue* (NASA CR-177354). Washington, D.C.: NASA.

Boggs, D. H., & Simon, J. R. (1968). Differential effect of noise on tasks of varying complexity. *Journal of Applied Psychology, 52*, 148-153.

Borg, C. G. (1978). Subjective aspects of physical and mental load. *Ergonomics, 21*, 215-220.

Bortolussi, M. R., Hart, S. G., & Shively, R. J. (1987). Measuring moment-to-moment pilot workload using synchronous presentations of secondary tasks in a motion-base trainer. In *Proceedings of the Fourth Symposium on Aviation Psychology.* Columbus, OH: Ohio State University.

Bortolussi, M. R., Kantowitz, B. H., & Hart, S. G. (1986). Measuring pilot workload in a motion base trainer. *Applied Ergonomics, 17,* 278-283.

Boyd, S. (1983). Assessing the validity of SWAT as a workload measurement instrument. In *Proceedings of Human Factors Society 27th Annual Meeting* (pp. 124-128). Santa Monica, CA: Human Factors Society.

Briggs, G. E., Peters, G. L., & Fisher, R. P. (1972). On the locus of the divided-attention effects. *Perception & Psychophysics, 11,* 315-320.

Broadbent, D. E., & Gregory, M. (1965). On the interaction of S-R compatibility with other variables affecting reaction time. *British Journal of Psychology, 56,* 61-67.

Broadbent, D. E., & Heron, A. (1962). Effects of a subsidiary task on performance involving immediate memory by younger and older men. *British Journal of Psychology, 53,* 189-198.

Brown, I. D. (1962). Measuring the "spare mental capacity" of car drivers by a subsidiary auditory task. *Ergonomics, 5,* 247-250.

Brown, I. D. (1965). A comparison of two subsidiary tasks used to measure fatigue in car drivers. *Ergonomics, 8,* 467-473.

Brown, I. D. (1966). Subjective and objective comparisons of successful and unsuccessful trainee drivers. *Ergonomics, 9,* 49-56.

Brown, I. D. (1967). Measurement of control skills, vigilance, and performance on a subsidiary task during twelve hours of car driving. *Ergonomics, 10,* 665-673.

Brown, I. D. (1968). Some alternative methods of predicting performance among professional drivers in training. *Ergonomics, 11,* 13-21.

Brown, I. D. (1982). Measurement of mental effort: Some theoretical and practical issues. In G. Harrison (Ed.), *Energy and effort* (pp. 27-37). London: Taylor and Francis.

Brown, I. D., & Poulton, E. C. (1961). Measuring the spare "mental capacity" of car drivers by a subsidiary task. *Ergonomics, 4,* 35-40.

Brown, I. D., Tickner, A. H., & Simmonds, D. C. V. (1969). Interference between concurrent tasks of driving and telephoning. *Journal of Applied Psychology, 53,* 419-424.

Brown, J. L. (1965). Flicker and intermittent stimulation. In C. H. Graham (Ed.), *Vision and visual perception.* New York: Wiley.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Card, S. K., Moran, T. P., & Newell, A. (1986). The model human processor: An engineering model of human performance. In K. R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of Perception and Human Performance: Vol. 2. Cognitive Processes and Performance* (pp. 45-1.-45-35.). New York: Wiley.

Casper, P. A., Shively, R. J., & Hart, S. G. (1987). Decision support for workload assessment: Introducing W C FIELDE. In *Proceedings of the Human Factors Society 31st Annual Meeting*. Santa Monica, CA: Human Factors Society.

Chapman, R. M. (1979). Connotative meaning and averaged evoked potentials. In H. Begleiter (Ed.), *Evoked brain potentials and behavior*. New York: Plenum Publishing Co.

Chechile, R. A., Butler, K., Gutowski, W., & Palmer, E. A. (1979). Division of attention as a function of the number of steps, visual shifts, and memory load. In *Proceedings of the 15th Annual Conference on Manual Control* (pp. 71-81). Dayton, OH: Wright State University.

Chiles, W. D., & Alluisi, E. A. (1979). On the specification of operator or occupational workload with performance-measurement methods. *Human Factors, 21*, 515-528.

Chiles, W. D., & Jennings, A. E. (1970). Effects of alcohol on complex performance. *Human Factors, 12*, 605-612.

Chiles, W. D., Jennings, A. E., & Alluisi, E. A. (1979). Measurement and scaling of workload in complex performance. *Aviation, Space and Environmental Medicine, 50*, 376-381.

Chow, S. L., & Murdock, B. B. (1975). The effect of a subsidiary task on iconic memory. *Memory and Cognition, 3*, 678-688.

Chubb, G. P., Laughery, K. R., & Pritsker, A. A. B. (1987). Simulating manned systems. In G. Salvendy (Ed.), *Handbook of Human Factors* (pp. 1298-1327). New York: John Wiley and Sons.

Comstock, E. M. (1973). Processing capacity in a letter-matching task. *Journal of Experimental Psychology, 100*, 63-72.

Cooper, G. E., & Harper, R. P. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (NASA TN-D-5153). Moffet Field, CA: NASA-Ames Research Center.

Courtright, J., & Kuperman, G. (1984). Use of SWAT in USAF system T & E. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 700-703). Santa Monica, CA: Human Factors Society.

Crabtree, M., Bateman, R., & Acton, W. (1984). Benefits of using objective and subjective workload measures. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 950-953). Santa Monica, CA: Human Factors Society.

Crawford, B. M. (1979). Workload assessment methodology development. In B. O. Hartman, & R. E. McKenzie (Eds.), *Survey of Methods to Assess Workload, AGARD-AG-246* (pp. 55-67).

Dalkey, N. C. (1969). *The Delphi Method: An experimental study of group opinion*. Santa Monica, CA: Rand Corporation.

Damos, D. L. (1978). Residual attention as a predictor of pilot performance. *Human Factors, 20*, 435-440.

Daniel, J., Florek, H., Kosinar, V., & Strizenec, M. (1969). Investigation of an operator's characteristics by means of factorial analysis. *Studia Psychologica, 11*, 10-22.

Darrow, C. M. (1929). Differences in the physiological reactions to sensory and ideational stimuli. *Psychological Bulletin, 26*, 185-201.

Derrick, W. L. (1983). Examination of workload measures with subjective task clusters. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 134-138). Santa Monica, CA: Human Factors Society.

Derrick, W. L. (1988). Dimensions of operator workload. *Human Factors, 30,* 95-110.

Detro, S. (1985). Subjective assessment of pilot workload in the advanced fighter cockpit. In *Proceedings of the Third Symposium on Aviation Psychology.* Columbus, OH: Ohio State University.

Dewar, R. E., Ellis, J. E., & Mundy, G. (1976). Reaction time as an index of traffic sign perception. *Human Factors, 18,* 381-392.

Dick, A. O. (1980). *Instrument scanning and controlling: Using eye movement data to understand pilot behavior and strategies* (CR-3306). Washington, DC: NASA.

Dick, A. O., Brown, J. L., & Bailey, G. (1978). Statistical evaluation of control inputs and eye movements in the use of instrument clusters during aircraft landing. NASA CR-149465. Washington, DC: NASA.

Dodge, R. (1903). Five types of eye movement in the horizontal meridian plane of the field of regard. *American Journal of Physiology, 8,* 307-329.

Donchin, E., Kramer, A. F., & Wickens, C. D. (1986). Applications of brain event-related potentials to problems in engineering psychology. In M.G.H. Coles, E. Donchin, & S. Porges (Eds.), *Psychophysiology: Systems, Processes, and Applications* (pp. 702-718). New York: Guilford Press.

Donchin, E., Ritter, W., & McCallum, W. C. (1978). Cognitive psychophysiology: The endogenous components of the ERP. In E. Callaway, P. Tueting, & S. Koslow (Eds.), *Event-Related Brain Potentials in Man* (pp. 349-441). New York: Academic Press.

Donnell, M. (1979). *An application of decision-analytic techniques to the test and evaluation phase of a major air system: Phase III* (TR-PR-79-6-91). McLean, VA: Decision and Designs, Inc.

Domic, S. (1980). Language dominance, spare capacity and perceived effort in bilinguals. *Ergonomics, 23,* 369-377.

Drury, C. G., Paramore, B., Van Cott, H. P., Grey, S. M., & Corlett, E. N. (1987). Task analysis. In G. Salvendy (Ed.), *Handbook of human factors.* New York: Wiley.

Duncan-Johnson, C. C. and Kopell, B. S. (1981). The stroop effect: Brain potentials localize the source of interference. *Science, 214,* 938-940.

Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: The variation in event-related potentials with subjective probability. *Psychophysiology, 14,* 456-467.

Duncan-Johnson, C. C., & Donchin, E. (1982). The P300 component of the event-related brain potential as an index of information processing. *Biological Psychology, 14,* 1- 52

Dunlap, W. P., Silver, N. C., & Bittner, A. C., Jr. (1986). Estimating reliability with small samples: Increased precision with averaged correlations. *Human Factors, 28,* 685-690.

Dyer, R., Matthews, J., Wright, C., & Yudowitch, K. (1976). *Questionnaire Construction Manual* (TCATA DAHC-19-74-C-0032). Ft. Hood, TX: ARI Special Publication P77-1. (AD A037 815)

Edwards, A. (1957). *Techniques of attitude construction.* New York: Appleton-Century-Crofts, Inc.

Edwards, R., Curnow, R., & Ostrand, R. (1977). *Workload assessment model (WAM) user's manual* (Report D180-20247-3). Seattle, WA: Boeing Aerospace Co.

Egelund, N. (1982). Spectral analysis of heart rate variability as an indicator of driver fatigue. *Ergonomics, 25*, 663-672.

Eggemeier, F. T., & Stadler, M. (1984). Subjective workload assessment in a spatial memory task. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 680-684). Santa Monica, CA: Human Factors

Eggemeier, F. T., Crabtree, M., & LaPointe, P. (1983). The effect of delayed report on subjective ratings of mental workload. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 139-143). Santa Monica, CA: Human Factors Society.

Eggemeier, F. T., Crabtree, M., Zingg, J., Reid, G., & Shingledecker, C. (1982). Subjective workload assessment in a memory update task. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 643-647). Santa Monica, CA: Human Factors Society.

Eggemeier, F. T., McGhee, J. Z., & Reid, G. (1983). The effects of variations in task loading on subjective workload rating scales. In *Proceedings of the IEEE 1983 National Aerospace and Electronics Conference* (pp. 1099-1106). Dayton, OH: IEEE.

Eggemeier, F. T., Melville, B., & Crabtree, M. (1984). The effect of intervening task performance on subjective workload ratings. In *Proceedings of the Human Factors Society 28th Annual Meetings* (pp. 954-958). Santa Monica, CA: Human Factors Society.

Eggleston, R. G. (1984). A comparison of projected and measured workload ratings using the Subjective Workload Assessment Technique (SWAT). In *Proceedings of the National Aerospace and Electronics Conference* (pp. 827-831). Dayton, OH: IEEE.

Eggleston, R. G., and Quinn, T. J. (1984). A preliminary evaluation of a projective workload assessment procedure. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 695-699). Santa Monica, CA: Human Factors Society.

Ellenbogen, B., & Dauley, R. (1962). Comparability of responses to a socially concordant question: "Open-end" and "Closed". *Journal of Health and Human Behavior, 3*(2), 136-140.

Ellis, J. E. (1973). Analysis of temporal and attentional aspects of movement control. *Journal of Experimental Psychology, 99*, 10-21.

Ellison, M. G., & Roberts, B. B. (1985). Timebased analysis of significant coordinated operations (TASCO): A cockpit workload analysis technique. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 774-778). Santa Monica, CA: Human Factors Society.

Enderwick, T. (1987). Human factors in operational test and evaluation. *Human Factors Society Test and Evaluation Technical Group Newsletter, 11*(1), 4-7.

England, L. (1948). Capital punishment and open-end questions. *Public Opinion Quarterly, 12*, 412-416.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251.

Ettema, J. H. (1969). Blood pressure changes during mental load experiments in man. *Psychother. Psychosom, 17*, 191-195.

Ettema, J. H., & Zielhuis, R. L. (1971). Physiological parameters of mental workload. *Ergonomics, 14*, 137-144.

Eysenck, M. (1977). *Human memory: Theory, research and individual differences.* New York: Pergamon.

Fadden, D. (1982). *Boeing Model 767 flight deck workload assessment methodology.* Paper presented at the SAE Guidance and Control System Meeting, Williamsburg, VA.

Farber, E., & Gallagher, V. (1972). Attentional demand as a measure of the influence of visibility conditions on driving task difficulty. *Highway Research Record, 414,* 1-5.

Figarola, T. R., & Billings, C. E. (1966). Effects of meprobamate and hypoxia on psychomotor performance. *Aerospace Medicine, 37,* 951-954.

Finkelman, J. M., & Glass, D. C. (1970). Reappraisal of the relationship between noise and human performance by means of a subsidiary task measure. *Journal of Applied Psychology, 54,* 211-213.

Fisher, S. (1975a). The microstructure of dual task interaction. 1. The patterning of main-task responses within secondary-task intervals. *Perception, 4,* 267-290.

Fisher, S. (1975b). The microstructure of dual task interaction. 2. The effect of task instructions on attentional allocation and a model of attentional-switching. *Perception, 4,* 459-474.

Fleishman, E. A. (1965). The prediction of total task performance from prior practice on task components. *Human Factors, 7,* 18-27.

Fournier, B. A., & Stager, P. (1976). Concurrent validation of a dual-task selection test. *Journal of Applied Psychology, 5,* 589-595.

Fruhstorfer, H., & Bergstrom, R. M. (1969). Human vigilance and auditory evoked potentials. *Electroencephalography and Clinical Neuro-physiology, 27,* 340-385

Gabay, E., & Merhav, S. J. (1977). Identification of a parametric model of the human operator in closed-loop control tasks. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-7,* 284-292.

Gabriel, R. F., & Burrows, A. A. (1968). Improving time-sharing performance of pilots through training. *Human Factors, 10,* 33-40.

Gainer, P. (1979). Analysis of visual estimation of system state from arbitrary displays. In M. C. Waller (Ed.), *Models of human operators in vision dependent tasks.* NASA Conference Publication 2103. Washington, DC: NASA.

Gale, A. (1977). Some EEG correlates of sustained attention. In R. R. Mackie (Ed.), *Vigilance: Theory, Operational Performance, and Physiological Correlates.* New York: Plenum.

Gale, A., Davies, R., & Smallbone, A. (1977). EEG Correlates of signal rate, time in task and individual differences in reaction time during a five-stage sustained attention task. *Ergonomics, 20,* 363-376.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts.* New York: Wiley.

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: Lawrence Erlbaum Associates.

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review, 63,* 317-329.

Gartner, W. B., & Murphy, M. R. (1976). *Pilot workload and fatigue: A critical survey of concepts and assessment techniques* (Report No. NASA-TN-D-8365). Washington, DC: NASA.

General Accounting Office (1982). *The Army needs to modify its system for measuring individual soldier proficiency* (GAO/FPCD-82-28). Washington, DC: Federal Personnel and Compensation Division. (NTIS AD-A112923)

Gescheider, G. (1985). *Psychophysics: Method, theory, and application.* Hillsdale, NJ: Lawerence Erlbaum Associates

Gidcumb, C. (1985). *Survey of SWAT use in flight test* (BDM/A-85-0630-TR). Albuquerque, NM: BDM Corporation.

Girouad, Y., Laurencelle, L., & Proteau, L. (1984). On the nature of the probe reaction-time task to uncover the attentional demands of movement. *Journal of Motor Behavior, 16*, 442-459.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5*, 351-360.

Goldstein, I. L., & Dorfman, P. W. (1978). Speed and load stress as determinants of performance in a time sharing task. *Human Factors, 20*, 603-609.

Gomer, F. E., Spicuzza, R. J., & O'Donnell, R. D. (1976). Evoked potentials correlates of visual item recognition during memory-scanning tasks. *Physiological Psychology, 4*, 61-65.

Gopher, D. (1977). Manipulating the conditions of training in time-share performance. *Human Factors, 19*, 583-593.

Gopher, D., & Braune, R (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors, 26*, 519-532.

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance. Vol. 2. Cognitive Processes and Performance* New York: John Wiley and Sons.

Gould, J. D., & Schaffer, A. (1967). The effects of divided attention on visual monitoring of mulitchannel displays. *Human Factors, 9*, 191-202.

Green, R., & Flux, R. Auditory communication and workload. In *Proceedings of the AGARD Conference on Methods to Assess Workload.* (AGARD-CP-216, pp A4-1 - A4-8).

Griffiths, I. D., & Boyce, P. R. (1971). Performance and thermal comfort. *Ergonomics, 14*, 457-468.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Hallett, P. E. (1986). Eye movements. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance. Vol. I. Sensory processes and perception.* New York: Wiley.

Hamilton, B. E., & Harper, H. P. (1984). Analytic methods for LHX mission and task analysis. In *Proceedings of Advanced Cockpit Specialist Meeting.* Washington, DC: American Helicopter Society.

Hamilton, P., Mulder, G., Strasser, H., & Ursin, H. (1979). Final report of physiological psychology group. In N Moray (Ed.), *Mental Workload, Its theory and measurement* (pp. 367-385). New York: Plenum Press.

Hansen, M. D. (1982). Keyboard design variables in dual-task. In *Proceedings of the 18th Annual Conference on Manual Control* (pp. 320-326). Dayton, OH: Flight Dynamics Laboratory.

Harris, R. L., Sr., & Christhilf, D. M. (1980). What do pilots see in displays? In *Proceedings of the Human Factors Society Meeting*. Santa Monica, CA: Human Factors Society.

Harris, R. L., Sr., & Glover, B. J. (1984). Effects of digital altimetry on pilot workload. Paper presented at the 1984 SAE Aerospace Congress and Exposition.

Harris, R. L., Sr., Glover, B. J., & Spady, A. A., Jr. (1986). *Analytic techniques of pilot scanning and their application.*(TP-2525). Washington, DC: NASA.

Harris, R. L., Sr., Tole, J. R., Stephens, A. T., & Ephrath, A. R. (1982). Visual scanning behavior and pilot workload. *Aviation, Space, and Environmental Medicine, 53*, 1067-1072.

Harris, R. M., Glenn, F., Iavecchia, H. P., & Zaklad, A. (1986). Human Operator Simulator. In W. Karwoski (Ed.), *Trends in ergonomic/human factors III (Part A)* (pp. 31-39). Amsterdam: North-Holland.

Harris, R. M., Iavecchia, H. P., Ross, L. V., & Shaffer, S. C. (1987). Microcomputer Human Operator Simulator (HOS-IV). In *Proceedings of the Human Factors Society 31st Annual Meeting*. Santa Monica, CA: Human Factors Society.

Hart, S. G. (1986a). Theory and measurement of human workload. In J. Zeidner (Ed.), *Human productivity enhancement. Vol. I* (pp 396-455). New York: Praeger.

Hart, S. G. (1986b). *Workload in complex systems.* Paper presented at the Symposium on the U. S. Army Key Operational Capabilities, The United States Army War College, Carlisle Barracks, PA.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock, & N. Meshkati (Eds.), *Human mental workload.* Amsterdam: Elsevier.

Hart, S. G., Battiste, V., & Lester, P. T. (1984). POPCORN: A supervisory control simulation for workload and performance research for workload and performance research (NASA-CP-2341). In *Proceedings of the 20th Annual Conference on Manual Control* (pp. 431-453). Washington, DC: NASA.

Hart, S. G., Childress, M. E., & Hauser, J. R. (1982). *Individual definitions of the term "workload."* In *Proceedings of the Eighth Symposium on Psychology in the Department of Defense* (pp. 478-485). Colorado Springs, CO: USAFA.

Hart, S. G., Shively, R. J., Vidulich, M. A., & Miller, R. C. (1985). The effects of stimulus modality and task integrality: Predicting dual-task performance and workload from single-task levels. In *Proceedings of the 21st Annual Conference on Manual Control* (pp. 5.1-5.18). Columbus, OH: NASA Ames Research Center and Ohio State University.

Hart. S. G. (1978). Subjective time estimation as an index of workload. In *Proceedings of the symposium on man-system interface: Advances in workload study* (pp. 115-131).

Harter, M. R., & Guido, W. (1980). Attention to pattern orientation: Negative cortical potentials, reaction time, and the selection process. *Electroencephalography and Clinical Neurophysiology, 49*, 461-475.

Harter, M. R., Previc, F. H., & Towle, V. L. (1979). Evoked potential indicants of size-and orientation-specific information processing: Feature-specific sensory channels and attention. In D. Lehmann, & E. Callaway (Eds.), *Human Evoked Potentials: Applications and Problems* (pp. 169-184). New York: Plenum.

Haworth, L., Bivens, C., & Shively, R. (1986). An investigation of single-piloted advanced cockpit and control configurations for nap-of-the-earth helicopter combat mission tasks. In *Proceedings of the 1986 Meeting of the American Helicopter Society* (pp. 657-672). Washington, DC: American Helicopter Society.

Hebb, D. O (1955). Drives and the C. N. S. (Conceptual nervous system). *Psychological Review, 62,* 243-254.

Heimstra, N. W. (1970). The effects of "stress fatigue" on performance in a simulated driving situation. *Ergonomics, 13,* 209-218.

Helm, W., & Donnell, M. (1979). *Mission operability assessment technique: A system evaluation methodology* (TP-79-31). Point Mugu, CA: Pacific Missile Test Center.

Helm, W., & Heimstra, N. (1981). *The relative efficiency of psychometric measures of task difficulty and task performance in predictive task performance* (Report No. HFL-81-5). Vermillion, SD: University of South Dakota, Psychology Department, Human Factors Laboratory.

Herman, L. M. (1965). Study of the single channel hypothesis and input regulation within a continuous, simultaneous task situation. *Quarterly Journal of Experimental Psychology, 17,* 37-46.

Hess, R. A. (1977). Prediction of pilot opinion ratings using an optimal pilot model. *Human Factors, 19,* 459-476.

Hess, R. A., & Teichgraber, W. M. (1974). Error quantization effects in compensatory tracking tasks. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-4,* 343-349.

Hicks, T. G., & Wierwille, W. W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors, 21,* 129-143.

Hilgendorf, E. L. (1967). Information processing practice and spare capacity. *Australian Journal of Psychology, 19,* 241-251.

Hill, S. G., Lysaght, R. J., Bittner, A. C., Jr., Bulger, J., Plamondon, B. D., Linton, P. M., & Dick, A. O. (1987). *Operator workload (OWL) assessment program for the Army: Results from requirements document reviews and user interview analysis* (DTR 2075-2). Willow Grove, PA: Analytics, Inc.

Hill, S. G., Plamondon, B. D., Wierwille, W. W., Lysaght, R. J., Dick, A. O., & Bittner, A. C., Jr. (1987). Analytic techniques for the assessment of operator workload. In *Proceedings of the Human Factors Society 31st Annual Meeting.* Santa Monica, CA: Human Factors Society.

Hill, S., & Bulger, J. (1988). Operator workload (OWL) in the Army materiel acquisition process.

Hillyard, S. A., & Kutas, M. (1983). Electrophysiology of cognitive processing. *Annual Review Psychology, 34,* 33-61.

Hoffman, E. R., & Jorbert, P. N. (1966). The effect of changes in some vehicle handling variables on driver steering performance. *Human Factors, 8,* 245-263.

Hoffman, J. E., Nelson, B., & Houck, M. R. (1983). The role of attentional resources in automatic detection. *Cognitive Psychology, 51,* 379-410.

Hohmuth, A. V. (1970). Vigilance performance in a bimodal task. *Journal of Applied Psychology, 54,* 520-525.

Holley, C. D., & Parks, R. E. (1987). Predicting man-machine system performance in predesign. Presented at the American Helicopter Society National Specialist Meeting on Flight Controls and Avionics, Cherry Hill, NJ.

Hollister, W. M. (Ed.). (1986). Allocation of human and automatic resources in the cockpit. In *Improved Guidance and Control Automation at the Man-Machine Interface, AGARD Advisory Report No. 228* (pp. 4-24). Neuilly Sur Seine, France: AGARD.

Howitt, J. S., Hay, A. E., Shergold, G. R., & Ferres, H. M. (1978). Workload and fatigue-in-flight EEG changes. *Aviation, Space and Environmental Medicine, 49,* 1196-1202.

Huddleston, J. H. F., & Wilson, R. V. (1971). An evaluation of the usefulness of four secondary tasks in assessing the effect of a lag in simulated aircraft dynamics. *Ergonomics, 14,* 371-380.

Isreal, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. (1980). P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. *Psychophysiology, 17,* 259-273.

Isreal, J. B., Wickens, C. D., Chesney, G. L., & Donchin, E. (1980). The event related brain potential as an index of display-monitoring workload. *Human Factors, 22,* 211-224.

Jahns, D. W. (1973). Operator workload: What is it and how should it be measured? In K. D. Cross, & J. J. McGrath (Eds.), *Crew System Design* (pp. 281-288). Santa Barbara, CA: Anacapa Sciences, Inc.

Jaschinski, W. (1982). Conditions of emergency lighting. *Ergonomics, 25,* 363-372.

Jobe, J. B., & Banderet, L. E. (1986). Cognitive testing in military performance research. In *Proceedings of a Workshop on Cognitive Testing Methodology*. Washington DC: National Research Council Committee on Military Nutrition Research.

Johannsen, G. (1979). Workload and workload measurement. In N. Moray (Ed.), *Mental workload, its theory and measurement* (pp. 3-11). New York: Plenum Press.

Johannsen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., & Wickens, C. (1979). Final report of experimental psychology group. In N. Moray (Ed.), *Mental Workload, its theory and measurement* (pp. 101-114). New York: Plenum Press.

John, P. G., Klein, G. A., & Taynor. J. (1986). Comparison-based prediction for front-end analysis. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 149-153). Santa Monica, CA: Human Factors Society.

Johnston, W. A., Greenberg, S. N., Fisher, R. P., & Martin, D. W. (1970). Divided attention: A vehicle for monitoring memory processes. *Journal of Experimental Psychology, 83,* 164-171.

Kahneman, D. (1973). *Attention and Effort.* Englewood Cliffs, N.J: Prentice-Hall.

Kahneman, D., Beatty, J., & Pollack, I. (1967). Perceptual deficit during a mental task. *Science, 157,* 218-219.

Kalsbeek, J. W. H. (1973). Do you believe in sinus arrhythmia? *Ergonomics, 16,* 99-104.

Kantowitz, B. H. (1985). Channels and stages in human information processing: A limited analysis of theory and methodology. *Journal of Mathematical Psychology, 29,* 135-174.

Kantowitz, B. H. (1987a). Mental workload. In P.A. Hancock (Ed.), *Human Factors Psychology.* Amsterdam: North Holland.

Kantowitz, B. H., & Knight, J. L. (1974). Testing tapping time-sharing. *Journal of Experimental Psychology, 103,* 331-336.

Kantowitz, B. H., & Knight, J. L. (1976). Testing tapping time-sharing: II. Auditory secondary task. *Acta Psychologica, 40,* 343-362.

Kantowitz, B. H., & Weldon, M. (1985). Scaling the axes of performance operating characteristic functions: Caveat emptor. *Human Factors, 27,* 531-547.

Keele, S. W. (1967). Compatibility and time-sharing in serial reaction time. *Journal of Experimental Psychology, 75,* 529-539.

Keele, S. W., & Boies, S. J. (1973). Processing demands of sequential information. *Memory and Cognition, 1,* 85-90.

Kelley, C. R. (1968). *Manual and automatic control.* New York: John Wiley and Sons.

Kelley, C. R., & Wargo, M. J. (1967). Cross-adaptive operator loading tasks. *Human Factors, 9,* 395-404.

Kelly, P. A., & Klapp, S. T. (1985). Hesitation in tracking induced by a concurrent manual task. In *Proceedings of the 21st Annual Conference on Manual Control* (pp. 19.1-19.3). Columbus, OH: Ohio State University.

Kirkpatrick, M., Malone, T. B., & Andrews, P. J. (1984). Development of an interactive microprocessor-based workload evaluation model (SIMWAM). In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 78-80). Santa Monica, CA: Human Factors Society.

Klapp, S. T., Kelly, P. A., Battiste, V., & Dunbar, S. (1984). Types of tracking errors induced by concurrent secondary manual task. In *Proceedings of the 20th Annual Conference on Manual Control* (pp. 299-304). Moffett Field, CA: Ames Research Center.

Klein, G. A. (1976). Effect of attentional demands on context utilization. *Journal of Educational Psychology, 68,* 25-31.

Knowles, W. B. (1963a). Operator loading tasks. *Human Factors, 5,* 155-161.

Knowles, W. B. (1963b). Operator loading tasks. *Human Factors, 6,* 357-383.

Kohen, S., de Mille, R., & Myers, J. (1972). Two comparisons of attitude measures. *Journal of Advertising Research, 12,* 29-34.

Kramer, A. F., Sirevaag, E. J., & Braune, R. (1987). A psychophysical assessment of operator workload during simulated flight missions. *Human Factors, 29,* 145-160.

Kramer, A. F., Wickens, C. D., & Donchin, E. (1984). Performance enhancements under dual-task conditions. In *Proceedings of the 20th Annual Conference on Manual Control* (pp. 21-35). Moffett Field, CA: Ames Research Center.

Kramer, A. F., Wickens, C. D., & Donchin, E. (1985). Processing of stimulus properties: Evidence for dual-task integrality. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 393-408.

Krantz, D., & Tversky, A. (1971). Conjoint-measurement analysis of composition rules in psychology. *Psychological Review, 78,* 151-169.

Kroese, B. S., & Siddle, D. A. T. (1983). Effects of an attention demanding task on the amplitude and habituation of the electrodermal orienting response. *Psychophysiology, 20,* 128-135.

Krol, J. P. (1971). Variations in ATC-workload as a function of variations in cockpit workload. *Ergonomics, 14*, 585-590.

Kuperman, G. G. (1985). Pro-SWAT applied to advanced helicopter crewstation concepts. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 398-402). Santa Monica, CA: Human Factors Society.

Kuperman, G. G., & Wilson, D. L. (1985). A workload analysis for strategic conventional standoff capability missions. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 635-639). Santa Monica, CA: Human Factors Society.

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory and Cognition, 11*, 539-550.

Kutas, M., McCarthy, G., & Donchin, E. (1977). Augmenting mental chronometry: The P300 as a measure of stimulus evaluation time. *Science, 197*, 792-795.

Kyriakides, K., & Leventhall, H. G. (1977). Some effects of infrasound on task performance. *Journal of Sound and Vibration, 50*, 369-386.

Lane, N. E. (1986). *Issues in performance measurement for military aviation with applications to air combat maneuvering* (NTSC TR-86-008). Orlando, FL: Naval Training Systems Center.

Lane, N. E., Kennedy, R. S., & Jones, M. B. (1986). Overcoming the unreliability of operational measures: The use of surrogate measure systems. In *Proceedings of the Human Factors 30th Annual Meeting* (pp.1398-1402). Santa Monica, CA: Human Factors Society.

Lane, N. E., Strieb, M. I., Glenn, F. A., & Wherry, R. J. (1981). The human operator simulator: An overview. In J. Moraal & K. F. Kraiss (Eds.), *Manned Systems Design: Methods, Equipment, and Applications* (pp. 121-152). New York: Plenum Press.

Lashley, K. S. (1929). *Brain mechanisms and intelligence*. Chicago, IL: The University of Chicago Press.

Laughery, K. R., Jr., Drews, C., Archer, R. & Kramme, K. (1986). A MicroSAINT simulation analyzing operator workload in a future attack helicopter. In *National Aerospace and Electronics Conference* (pp. 896-903). Dayton, OH: IEEE.

Laughery, K. R., Sr., & Laughery, K. R., Jr. (1987). Analytic techniques for function analysis. In G. Salvendy (Ed.), *Handbook of human factors*. New York: Wiley.

Laurell, H., & Lisper, H. O. (1978). A validation of subsidiary reaction time against detection of roadside obstacles during prolonged driving. *Ergonomics, 21*, 81-88.

Levison, W. H. (1970). A model for task interference. In *Proceedings of the 6th Annual Conference on Manual Control* (pp. 585-616). Wright-Patterson AFB, OH.

Levison, W. H. (1979). A model for mental workload in tasks requiring continuous information processing. In N. Moray (Ed.), *Mental Workload: Its theory and measurement* (pp. 189-218). New York: Plenum Press.

Levison, W. H., & Tanner, R. B. (1971). *A control-theory model for human decision making*. NASA CR-1953. Washington, DC: NASA.

Lewis, R. E. F., Dela Riviere, W. D., & Sweeney, D. M. (1968). Dual versus solo pilot navigation in helicopters at low level. *Ergonomics, 11*, 145-155.

Lidderdale, I. G. (1987). Measurement of aircrew workload during low-level flight . In A. H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp. 69-77). Neuilly Sur Seine, France: AGARD.

Lidderdale, I. G., & King, A. H. (1985). *Analysis of subjective ratings using the analytical hierarchy process; A microcomputer program.* High Wycombe, England: OR Branch NFR, HQ STC, RAF.

Lindsay, P. H., & Norman, D. A. (1969). Short-term retention during a simultaneous detection task. *Perception and Psychophysics, 5,* 201-205.

Linton, P. M., Jahns, D. W., & Chatelier, P. R. (1977). Operator workload assessment model: An evaluation of a VF/VA-V/STOL system. In *Proceedings of the AGARD Conference on Methods to Assess Workload* (AGARD-CP-216, pp. A12-1 - A12-11).

Lisper, H. O., Laurell, H., & Stening, G. (1973). Effects of experience of the driver on heart-rate, respiration-rate, and subsidiary reaction time in a three hours continuous driving task. *Ergonomics, 16,* 501-506.

Liu, Y-Y., & Wickens, C. D. (1987). *Mental workload and cognitive task automation: An evaluation of subjective and time estimation metrics.* Tech Report No. EPL-87-21, NASA 87-2. Champaign, IL: University of Illinois Aviation Research Laboratory.

Logan, G. D. (1970). On the use of a concurrent memory load to measure attention and automaticity. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 189-207.

Long, J. (1976). Effect of task difficulty on the division of attention between nonverbal signals: independence or interaction? *Quarterly Journal of Experimental Psychology, 28,* 179-192.

Looper, M. (1976). The effect of attention loading on the inhibition of choice reaction time to visual motion by concurrent rotary motion. *Perception and Psychophysics, 20,* 80-84.

Luria, A. R. (1966). *Higher cortical functions in man.* New York: Basic Books.

Malmo, R. B. (1959). Activation: A neurophysiological dimension. *Psychological Review, 66,* 367-386.

Malmstrom, F. V., Reed, L. E., & Randle, R. J. (1983). Restriction of pursuit eye movement range during a concurrent auditory task. *Journal of Applied Psychology, 68,* 565-571.

Malone, T. B., Kirkpatrick, M., & Kopp, W. H. (1986). Human factors engineering impact of system workload and manning levels. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 763-767). Santa Monica, CA: Human Factors Society.

Mandler, G., & Worden, P. E. (1973). Semantic processing without permanent storage. *Journal of Experimental Psychology, 100,* 277-283.

Martin, D. W., & Kelly, R. T. (1974). Secondary task performance during directed forgetting. *Journal of Experimental Psychology, 103,* 1074-1079.

Masline, P. J. (1986). *A comparison of the sensitivity of interval scale psychometric techniques in the assessment of subjective mental workload.* Unpublished masters thesis, University of Dayton, Dayton, OH.

McCarthy, G., & Donchin, E. (1981). A metric for thought: A comparison of P300 latency and reaction time. *Science, 211,* 77-80.

McCracken, J. H., & Aldrich, T. B. (1984). *Analysis of selected LHX mission functions: Implications for operator workload and system automation goals* (TNA ASI479-24-84). Fort Rucker, AL: Anacapa Sciences, Inc.

McGrath, J. J. (1965). Performance sharing in an audio-visual vigilance task. *Human Factors, 7,* 141-153.

McLeod, P. D. (1973). Interference of "attend to and learn" tasks with tracking. *Journal of Experimental Psychology, 99,* 330-333.

Meister, D. (1985). *Behavioral analysis and measurement methods.* New York: John Wiley and Sons.

Meister, D. (1986). A survey of test and evaluation practices. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 1239-1243). Santa Monica, CA: Human Factor Society.

Mewhort, D. J. K., Thio, H., & Birkmayer, A. C. (1971). Processing capacity and switching attention in dichotic listening. *Canadian Journal of Psychology, 25,* 111-129.

Michon, J. A. (1964). A note on the measurement of perceptual motor load. *Ergonomics, 7,* 461-463.

Michon, J. A. (1966). Tapping regularity as a measure of perceptual motor load. *Ergonomics, 9,* 401-412.

Miles, W. R. (Ed.) (1936). Psychological studies of human variability. *Psychological Monographs, 67,* Whole No. 212.

Miller, K. (1975). Processing capacity requirements of stimulus encoding. *Acta Psychologica, 39,* 393-410.

Miller, R. C., & Hart, S. G. (1984). Assessing the subjective workload of directional orientation tasks (NASA-CP-2341). In *Proceedings of the 20th Annual Conference on Manual Control* (pp. 85-95). Washington, DC.: NASA.

Mirchandani, P. B. (1972). An auditory display in a dual axis tracking task. *IEEE Transactions on Systems, Man, and Cybernetics, 2,* 375-380.

Mitsuda, M. (1968). Effects of a subsidiary task on backward recall. *Journal of Verbal Learning and Verbal Behavior, 7,* 722-725.

Monty, R. A., & Ruby, W. J. (1965). Effects of added workload on compensatory tracking for maximum terrain following. *Human Factors, 7,* 207-214.

Moray, N. (1976). Attention, control and sampling behavior. In T. B. Sheridan, & G. Johannsen, (Eds.), *Monitoring behavior and supervisory control* (pp. 221-244). New York: Plenum Press.

Moray, N. (1979a). Models and measures of mental workload. In N. Moray (Ed.), *Mental Workload: Its theory and measurement* (pp.13-21). New York: Plenum Press.

Moray, N. (1982). Subjective mental workload. *Human Factors, 24,* 25-40.

Moray, N. (Ed.). (1979b). *Mental workload: Its theory and measurement.* New York: Plenum.

Moses, R. A. (1970). *Adler's Physiology of the eye. Clinical application.* St. Louis: C.V. Mosby.

Moskowitz, H., & McGlothlin, W. (1974). Effects of marihuana on auditory signal detection. *Psychopharmacologia, 40,* 137-145.

Mulder, I. J. M., & Mulder, G. (1987). Cardiovascular reactivity and mental workload. In R. I. Kitney & O. Rompleman (Eds.), *The beat-to-beat investigation of cardiovascular function.* New York: Oxford.

Murdock, B. B. (1965). Effects of a subsidiary task on short-term memory. *British Journal of Psychology, 56*, 413-419.

NASA-Ames Research Center, Human Performance Group (1986). *Collecting NASA workload ratings: A paper-and-pencil package* (Version 2.1). Moffet Field, CA: NASA-Ames Research Center.

Natani, K., & Gomer, F. E. (1981). *Electrocortical activity and operator workload: A comparison of changes in the electroencephalogram and in event-related potentials.* McDonnell Douglas Tech, Report E2477. St. Louis, MO. McDonnell Douglas Astronautics Company

Navon, D. (1984). Resources – A theoretical soup stone? *Psychological Review, 91*, 216-234.

Navon, D., & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review, 86*, 214-255.

Nideffer, R. M. (1976). *The inner athlete.* New York: Thomas Y. Crowell Company.

Noble, C. E. (1961). Verbal learning and individual differences. In C. N. Cofer (Ed.), *Verbal learning and verbal behavior.* New York: McGraw-Hill.

Noble, M., Trumbo, D., & Fowler, F. (1967). Further evidence on secondary task interference in tracking. *Journal of Experimental Psychology, 73*, 146-149.

Norman, D., & Bobrow, D. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*, 44-64.

North, R. A. (1986). A workload index for iterative crewstation evaluation. In *Proceedings of the Eighth Annual Carmel Workshop: Workload and Training, an Examination of Their Interactions.*

North, R. A., Stackhouse, S. P., & Graffunder, K. (1979). *Performance, physiological and oculometer evaluation of VTDL landing displays* (NASA Contractor Report 3171). Hampton, VA: NASA-Langley Research Center.

Notestine, J. (1984). Subjective workload assessment and effect of delayed ratings in a probability monitoring task. In *Proceedings of the Human Factors Society 28th Annual Meeting* (pp. 685-690). Santa Monica, CA: Human Factors Society.

Nygren, T. E. (1982). *Conjoint measurement and conjoint scaling: A user's guide.* (AFAMRL-TR-82-22). Wright-Patterson AFB, OH: AFAMRL.

Nygren, T. E. (1985). Axiomatic and numeric conjoint measurement: A comparison of three methods for obtaining subjective workload (SWAT) ranking. In *Proceedings of the National Aerospace and Electronics Conference.* Dayton, OH: IEEE.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance. Vol. 2. Cognitive Processes and Performance* New York: John Wiley and Sons.

O'Hanlon, J., & Beatty, J. (1977). Concurrence of EEG and performance changes during a simulated radar watch and some implications for the arousal theory of vigilance. In R.R. Mackie (Ed.), *Vigilance: Theory, Operational Performance, an Physiological Correlates.* New York: Plenum.

Ogden, G. D., Levine, J. M., & Eisner, E. J. (1979). Measurement of workload by secondary tasks. *Human Factors, 21*, 529-548.

Oshima, M. (1981). The mechanism of mental fatigue. In G. Salvendy & M. J. Smith (Eds.), *Machine pacing and occupational stress.* London: Taylor & Francis.

Ozier, M. (1980). Individual differences in free recall. In G. H. Bower (Ed.), *The psychology of learning and motivation.* New York: Academic Press.

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition.* Hillsdale, NJ: Erlbaum Associates.

Parasuraman, R (1986). Vigilance, monitoring and search. In K. R. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance. Vol. 2. Cognitive Processes and Performance* New York: John Wiley and Sons.

Parasuraman, R. (1984). Sustained attention in detection and discrimination. In R. Parasuraman, & D. R. Davies (Eds.), *Varieties of Attention* (pp. 243-271). New York: Academic Press.

Peterson, N. G. (1985). Overall strategy and methods for expanding the measured predictor space. Symposium: Expanding the measurement of predictor space for military enlisted jobs. Presented at the Annual meeting of the American Psychological Association, Los Angles, CA.

Pew, R. W. (1974). Human perceptual-motor performance. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Posner, M. I. (1978). *Chronometric explorations of mind.* Hillsdale, NJ: Erlbaum.

Posner, M. I. (1986). Overview. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance. Vol. 2. Cognitive Processes and Performance* New York: John Wiley and Sons.

Potter, S. (1986). *Subjective workload assessment technique (SWAT) subscale sensitivity to variations in task demand and presentation rate.* Unpublished masters thesis, Wright State University, Dayton, OH.

Potter, S., & Acton, W. (1985). Relative contributions of SWAT dimensions to overall subjective workload ratings. In *Proceedings of Third Symposium on Aviation Psychology* Columbus, OH: Ohio State University.

Price, D. L. (1975). The effects of certain gimbal orders on target acquisition and workload. *Human Factors, 17,* 571-576.

Prien, E., Otis, J., Campbell, J., & Saleh, S. (1964). Comparison of methods of measurement of job attitudes. *Journal of Industrial Psychology, 2,* 87-97.

Pritchard, W. S. (1981). Psychophysiology of P300. *Psychological Bulletin, 89,* 506-540.

Putz, V. R., & Rothe, R. (1974). Peripheral signal detection and concurrent compensatory tracking. *Journal of Motor Behavior, 6,* 155-163.

Rault, A (1976). Pilot workload analysis. . In T. B. Sheridan & G. Johannsen (Eds.), *Monitoring Behavior and Supervisory Control.* (pp.139-155). New York: Plenum Press.

Regan, D. (1977). Steady-state evoked potentials. *Journal of the Optical Society of America, 67,* 1475-1489.

Reid, G. B., Eggemeier, F. T., & Shingledecker, C. A. (1983). *Workload analysis for the AMRAAM operational test and evaluation.* Wright-Patterson AFB, OH: Air Force Aerospace Medical Research Laboratory.

Reid, G. B., Eggemeier, F., & Nygren, T. (1982). An individual differences approach to SWAT scale development. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 639-642). Santa Monica, CA: Human Factors Society.

Reid, G. B., Shingledecker, C. A, Hockenberger, R., & Quinn, T. J. (1984). A projective application of the subjective workload assessment techniques. In *Proceedings of the National Aerospace and Electronics Conference* (pp. 824-826). Dayton, OH: IEEE.

Reid, G. B., Shingledecker, C. A., & Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. In *Proceedings of the Human Factors Society 25th Annual Meeting* (pp. 522-525). Santa Monica, CA: Human Factors Society.

Rickard, W. W., & Levison, W. H. (1981). Further tests of a model-based scheme for predicting pilot opinion ratings for large commercial transports. In *Proceedings of the 17th Annual NASA-University Conference on Manual Control* (pp. 245-254). University of California at Los Angeles.

Roberts, B. B., & Crites, C. D. (1985). Computer-aided cockpit workload analysis for all weather, multirole tactical aircraft. In *Fourth Aerospace Behavioral Engineering Technology Conference Proceedings, SAE Paper 851876* (pp. 111-123). Warrendale, PA: Society of Automotive Engineers.

Roediger, H. L., Knight, J. L., & Kantowitz, B. H. (1977). Inferring decay in short-term memory: The issue of capacity. *Memory and Cognition, 5,* 167-176.

Rohmert, W. (1987). Physiological and psychological workload measurement and analysis. In G. Salvendy (Ed.), *Handbook of human factors* (pp. 402-428). New York: John Wiley and Sons.

Rolfe, J. M. (1971). The secondary task as a measure of mental load. In W.T. Singleton, J.G. Fox, & D. Whitfield (Eds.), *Measurement of man at work* (pp. 135-148). London: Taylor and Francis.

Roscoe, A. H (1987a). In-flight assessment of workload using pilot ratings and heart rate. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp. 78-82). Neuilly Sur Seine, France: AGARD.

Roth, W. T., Ford, J. M., & Kopell, B. S. (1978). Long-latency evoked potentials and reaction time. *Psychophysiology, 15,* 17-23.

Rouse, W. B. (1980). *Systems Engineering Models of Human-Machine Interaction.* New York: Elsevier North Holland.

Ruggiero, F., & Fadden, D. (1987). Pilot subjective evaluation of workload during a flight test certification programme. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload* AGARDograph 282 (pp. 32-36). Neuilly Sur Seine, France: AGARD.

Saaty, T.L. (1980). *The analytical hierarchy process.* New York: McGraw-Hill.

Sanders, A. F. (1979). Some remarks on mental load. In N. Moray (Ed.), *Mental workload: Its theory and measurement* (pp. 41-78). New York: Plenum Press.

Sanders, M. G., Burden, R. T., Jr., Simmons, R. R., Lees, M. A., & Kimball, K. A. (1978). An evaluation of perceptual-motor workload during a helicopter hover maneuver. USAARL Report 78-14. Ft. Rucker, AL: U. S. Army Aeromedical Research Laboratory.

Savage, R. E., Wierwille, W. W., & Cordes, R. E. (1978). Evaluating the sensitivity of various measures of operator workload using random digits as a secondary task. *Human Factors, 20,* 649-654.

Sayers, B. (1973). Analysis of heart rate variablity. *Ergonomics, 16,* 17-32.

Scales, D., & Yeoman, A. (1950). *Developing an objective item questionnaire to assess the market for further education among employed adults.* Washington, DC: American Council on Education.

Schick, F.V., & Hann, R. L. (1987). The use of subjective workload assessment technique in a complex flight task. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp. 37-41). Neuilly Sur Seine, France: AGARD.

Schiflett, S. G. (1976). *Operator workload: An annotated bibliography* (SY-257R-76). Patuxent River, MD: U.S. Navy Air Test Center.

Schmidt, D. K. (1978). A queueing analysis of the air traffic controller's workload. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-8*, 492-493.

Schmidt, K. H., Kleinbeck, U., & Brockmann, W. (1984). Motivational control of motor performance by goal setting in a dual-task situation. *Psychological Research, 46*, 129-141.

Schneider, W., & Fisk, A. D. (1982). Degree of consistent training: Improvements in search performance and automatic process development. *Perception and Psychophysics, 31*, 160-168.

Schori, T. R. (1973). A comparison of visual, auditory, and cutaneous tracking displays when divided attention is required to a cross-adaptive loading task. *Ergonomics, 16*, 153-158.

Schori, T.R., & Jones, B.W. (1975). Smoking and work load. *Journal of Motor Behavior, 7*, 113-120.

Schouten, J. F., Kalsbeek, J. W. H., & Leopold, F.F. (1962). On the evaluation of perceptual and mental load. *Ergonomics, 5*, 251-260.

Schvaneveldt, R. W. (1969). Effects of complexity in simultaneous reaction time tasks. *Journal of Experiment Psychology, 81*, 289-296.

Senders, J. W. (1964). The human operator as a monitor and controller of multi-degree of freedom systems. *IEEE Transactions on Human Factors and Electronics, HFE-5*, 2-5.

Senders, J. W. (1979). Axiomatic levels of workload. In N. Moray (Ed.), *Mental Workload: Its Theory and Measurement* (pp. 263-267). New York: Plenum Press.

Senders, J. W. Elkind, J. I., Grignetti, M. C., & Smallwood, R. 1966). *An investigation of the visual sampling behavior of human observers.* Bolt, Beranek and Newman, Inc. NASA CR-434. Washington, DC: NASA.

Senders, J. W., & Posner, M. (1976). A queueing model of monitoring and supervisory behavior. In T. B. Sheridan, & G. Johannsen, (Eds.), *Monitoring Behavior and Supervisory Control* (pp. 245-259). New York: Plenum Press.

Senders, J. W., Kristofferson, A. B., Levison, W. H., Dietrich, C. W., & Ward, J. L. (1967). The attentional demand of automobile driving. *Highway Research Record, 195*, 15-33.

Shaffer, M. T., Shafer, J. B., & Kutch, G. B. (1986). Empirical workload and communication: Analysis of scout helicopter exercises. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 628-632). Santa Monica, CA: Human Factors Society.

Sharit, J., & Salvendy, G. (1982). Occupational stress: Review and reappraisal. Human Factors, 24, 129-162.

Sheridan, T. B. (1980). Mental workload - What is it? Why bother with it? *Human Factors Society Bulletin, 23*, 1-2.

Sheridan, T. B., & Ferrell, W. R. (1974). *Man-Machine Systems: Information, Control, and Decision Models of Human Performance.* Cambridge, MA: MIT Press.

Sheridan, T. B., & Simpson, R. W. (1979). *Toward the definition and measurement of the mental workload of transport pilots* (FTL Report R79-4). Cambridge, MA: Flight Transportation Laboratory.

Shingledecker, C. A. (1983, April). Behavior and subjective workload metrics for operational environments. In *Proceedings of the AGARD (AMP) Symposium on Sustained Intensive Air Operations: Physiological and Performance Aspects.* Paris, France: AGARD.

Shingledecker, C. A. (1987). In-flight workload assessment using embedded secondary radio communications tasks. In A. H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp. 11-14). Neuilly Sur Seine, France: AGARD.

Shingledecker, C. A., & Crabtree, M. S. (1982). *Subsidiary radio communications tasks for workload assessment in R&D simulations: II. Task sensitivity evaluation* (AFAMRL-TR-82-57). Wright-Patterson AFB, OH: Aerospace Medical Research Laboratory.

Shingledecker, C. A., Acton, W. H., & Crabtree, M. S. (1983). *Development and application of a criterion task set for workload metric evaluation* (SAE Technical Paper No. 831419). Warrendale, PA: Society of Automotive Engineers.

Shingledecker, C. A., Crabtree, M. S., Simons, J. C., Courtright, J. F., & O'Donnell, R. D. (1980). *Subsidiary radio communications tasks for workload assessment in R&D simulations: I. Task development and workload scaling* (AFAMRL-TR-80-126). Wright-Patterson AFB, OH: Aerospace Medical Research Laboratory.

Shively, R., Battiste, V., Matsumoto, J., Pepitone, D., Bortolussi, M., & Hart, S. G. (1987). Inflight evaluation of pilot workload measures for rotorcraft research. In *Proceedings of the Fourth Symposium on Aviation Psychology.* Columbus, OH: Ohio State University.

Shulman, H. G., & Greenberg, S. N. (1971). Perceptual deficit due to division of attention between memory and perception. *Journal of Experimental Psychology, 88*, 171-176.

Shulman, H. G., Greenberg, S. N., & Martin, J. (1971). Intertask delay as a parameter of perceptual deficit in divided attention. *Journal of Experimental Psychology, 88*, 439-440.

Siegel, A. I., & Wolf, J. J. (1969). *Man-Machine Simulation Models: Psychosocial and Performance Interaction.* New York: John Wiley and Sons.

Silverstein, C., & Glanzer, M. (1971). Concurrent task in free recall: Differential effects of LTS and STS. *Psychonomic Science, 22*, 367-368.

Skelly, J., & Purvis, B. (1985, April). *B-52 wartime mission simulation: Scientific precision in workload assessment.* Paper presented at the 1985 Air Force Conference on Technology in Training and Education, Colorado Springs, CO.

Smit, J., & Wewerinke, P. H. (1978, May). *An analysis of helicopter pilot control behavior and workload during instrument flying tasks.* Paper presented at the AGARD Aerospace Medical Panel Specialists Meeting on Operational Helicopter Aviation Medicine.

Smith, M. C. (1969). Effect of varying channel capacity on stimulus detection and discrimination. *Journal of Experimental Psychology, 82*, 520-526.

Smith, R. L., Lucaccini, L. F., Groth, H., & Lyman, J. (1966). Effects of anticipatory alerting signals and a compatible secondary task on vigilance performance. *Journal of Applied Psychology, 50*, 240-246.

Soliday, S. M, & Schohan, B. (1965). Task loading of pilots in simulated low-altitude high-speed flight. *Human Factors, 7*, 45-53.

Spady, A. A., Jr. (1978a). *Airline pilot scan patterns during simulated ILS approaches.* NASA TP-1250. Washington, DC: NASA

Spady, A. A., Jr. (1978b). Airline scanning behavior during approaches and landing in a Boeing 737 simulator. *Guidance and Control Design Considerations for Low-Altitude and Terminal-Area Flight*, AGARD-CP-240, 17-1-17-5.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Speyer, J., Fort, A., Fouillot, J., & Blomberg, R. (1987). Assessing pilot workload for minimum crew certification. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp. 90-115). Neuilly Sur Seine, France: AGARD.

Spicuzza, R., Pincus, A., & O'Donnell, R. D. (1974). *Development of performance assessment methodology for the digital avionics information system* (Technical report). Dayton, OH: Systems Research Laboratories, Inc.

Stager, P., & Muter, P. (1971). Instructions and information processing in a complex task. *Journal of Experimental Psychology, 87*, 291-294.

Stager, P., & Zufelt, K. (1972). Dual-task method in determining load differences. *Journal of Experimental Psychology, 94*, 113-115.

Sternberg, S. (1966). High speed scanning in human memory. *Science, 153*, 652-654.

Sternberg, S. (1969). Memory scanning: Mental processes revealed by reaction time experiments. *American Scientist, 57*, 421-457.

Stevens, S. S (1975). *Psychophysics.* New York: Wiley.

Stevens, S. S. (1958). Problems and methods of psychophysics. *Psychological Bulletin, 55*, 177-196.

Sticha, P. J. (1987). Models of procedural control for human performance simulation. *Human Factors, 29*, 421-432.

Stone, G., Gulick, R. K., & Gabriel, R. F. (1987). Use of timeline analysis to assess crew workload. In A. H. Roscoe, (Ed.), *The Practical Assessment of Pilot Workload, AGARD-AG-282* (pp. 15-31). Neuilly Sur Seine, France: AGARD.

Strasser, H. (1981). Physiological measures of mental load. In E. N. Corlett & J. Richardson (Eds.). *Stress, work design, and productivity.* New York: Wiley.

Strasser, H. (1985). Assessment of psychomental workload in modern factories. In K. Naro (Ed.), *Occupational health and safety in automation and robotics.* (The Proceedings of the 5th UOEH International Symposium.) London: Taylor & Francis.

Strasser, H. (1987). Assessment of psychomental workload in modern factories. In K. Noro (Ed.), *Occupational health and safety in automation and robotics* (pp. 288-308). London: Taylor and Francis.

Sutton, S., Braren, M. Zubin, J., & John, E. R. (1965). Evoked potential correlates of stimulus uncertainty. *Science, 150,* 1187-1188.

Szabo, S. M., Bierbaum, C. R., & Hocutt, C. D. (1987). Development and validation of a workload prediction methodology for AH-64 crewmembers. Paper presented at DoD Technical Advisory Group, 16-19 November, Oxnard, CA.

Thompson, M. W., & Bateman, R. P. (1986). *A computer-based workload prediction model.* SAE AeroTech.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometr. Monographs,* No. 1.

Tickner, A. H., Poulton, E.C., Copeman, A. K., & Simmonds, D. C. V. (1972). Monitoring 16 television screens showing little movement. *Ergonomics, 15,* 279-291.

Tole, J. R., Stephens, A. T., Harris, R. L., Sr., & Ephrath, A. R. (1982). Visual scanning behavior and mental workload in aircraft pilots. *Aviation, Space, and Environmental Medicine, 53,* 54-61.

Truijens, C. L., Trumbo, D. A., & Wagenaar, W. A. (1976). Amphetamine and barbituate effects on two tasks performed singly and in combination. *Acta Psychologica, 40,* 233-244.

Trumbo, D., & Milone, F. (1971). Primary task performance as a function of encoding, retention, and recall in a secondary task. *Journal of Experimental Psychology, 91,* 273-279.

Trumbo, D., & Noble, M. (1970). Secondary task effects on serial verbal learning. *Journal of Experimental Psychology, 85,* 418-424.

Trumbo, D., & Noble, M. (1972). Response uncertainty in dual-task performance. *Organizational Behavior and Human Performance, 7,* 203-215.

Trumbo, D., Noble, M., & Swink, J. (1967). Secondary task interference in the performance of tracking tasks. *Journal of Experimental Psychology, 73,* 232-240.

Tsang, P. S., & Johnson, W. (1987). Automation: Changes in cognitive demands and mental workload. In *Proceedings of the Fourth Symposium on Aviation Psychology.* Columbus, OH: Ohio State University.

Tsang, P. S., & Wickens, C. D. (1984). The effects of task structures on time-sharing efficiency and resource allocation optimality. In *Proceedings of the 20th Annual Conference on Manual Control* (pp. 305-317). Moffett Field, CA: Ames Research Center.

Tyler, D. M., & Halcomb, C. G. (1974). Monitoring performance with a time-shared encoding task. *Perceptual and Motor Skills, 38,* 383-386.

U. S. Army (1987). *Manpower and Personnel Integration (MANPRINT) in Materiel Acquisition Process* (AR 602-2). Washington, DC: Department of the Army.

U. S. Army Soldier Support Center (1986). *Early Comparability Analysis (ECA) Procedural Guide.* Washington, DC: Department of the Army.

U.S. Army (1979). *Military Specification: Human Engineering Requirements for Military Systems, Equipment and Facilities* (MIL-H-46855B). Washington, DC: Department of the Army.

U.S. Army (1983). Human Factors Engineering Program (AR 602-1). Washington, DC: Department of the Army.

U.S. Army Test and Evaluation Command (1976). *Questionnaire and interview design, Subjective testing techniques* (TECOM Pam 602-1, Vol. I). Aberdeen Proving Ground: USATECOM.

Van Horn, M. (1986). *Understanding Expert Systems.* New York: The Waite Group, Inc.

Vidulich, M. A. (1988). The cognitive psychology of subjective mental workload. In P. A. Hancock, & N. Meshkati (Eds.), *Human mental workload.* Amsterdam, The Netherlands: Elsevier.

Vidulich, M. A., & Pandit, P. (1986). Training and subjective workload in a category search task. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 1133-1136). Santa Monica, CA: Human Factors Society.

Vidulich, M. A., & Tsang, P. S. (1985a). Assessing subjective workload assessment: A comparison of SWAT and the NASA-Bipolar methods. In *Proceedings of the Human Factors Society 29th Annual Meeting* (pp. 71-75). Santa Monica, CA: Human Factors Society.

Vidulich, M. A., & Tsang, P. S. (1985b). Techniques of subjective workload assessment: A comparison of two methodologies. In R. Jensen, & J. Adrion (Eds.), *Proceedings of the Third Symposium on Aviation Psychology* (pp. 239-246). Columbus, OH: OSU Aviation Psychology Laboratory.

Vidulich, M. A., & Tsang, P. S. (1985c). Evaluation of two cognitive abilities tests in a dual-task environment. In *Proceedings of the 21st Annual Conference on Manual Control.* (pp. 12.1-12.10). Columbus, OH: Ohio State University.

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. In *Proceedings of the Human Factors Society 31st Annual Meeting.* (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Vidulich, M. A., & Wickens, C. (1986). Causes of dissociation between subjective workload measures and performance: Caveats of the use of subjective assessments. *Applied Ergonomics, 17,* 291-296.

Vincente, K. J., Thornton, D. C., & Moray, N. (1987). Spectral analysis of sinus arrhythmia: A measure of mental effort. *Human Factors, 29,* 171-182.

Vroon, P. A. (1973). Tapping rate as a measure of expectancy in terms of response and attention limitation. *Journal of Experimental Psychology, 101,* 183-185.

Wainwright, W. (1987). Flight test evaluation of crew workload. In A.H. Roscoe (Ed.), *The practical assessment of pilot workload, AGARDograph No. 282* (pp. 60-68). Neuilly Sur Seine, France: AGARD.

Walden, R. S., & Rouse, W. B. (1978). A queueing model of pilot decision making in a multi-task flight management situation. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-8,* 867-875.

Waller, M. C. (1976). *An investigation of correlation between pilot scanning behavior and workload using stepwise regression analysis* (TM X-3344). Washington, DC: NASA.

Warr, D., Colle, H., & Reid, G. (1986). A comparative evaluation of two subjective workload measures: The Subjective Workload Assessment Technique and the Modified Cooper Harper Scale. Paper presented at the *Symposium on Psychology in the Department of Defense,* USAFA, Colorado Springs, CO.

Wastell, D. G, & Kleinman, D. (1980). Evoked potential correlates of visual selective attention. *Acta Psychologica, 46,* 129-140.

Watson, A. B. (1986). Temporal sensitivity. In In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance. Vol. I. Sensory processes and perception.* New York: Wiley.

Weichselgartner, E. & Sperling, G. (1987). Dynamics of automatic and controlled visual attention. *Science, 238,* 778-780.

Wempe, T. E., & Baty, D. L. (1968). Human information processing rates during certain multiaxis tracking tasks with a concurrent auditory task. *IEEE Transactions On Man-Machine System, 9,* 129-138.

Wetherell, A. (1981). The efficacy of some auditory-vocal subsidiary tasks as measures of the mental load on male and female drivers. *Ergonomics, 24,* 197-214.

Wewerinke, P. H. (1974). Human operator workload for various control situations. *Tenth Annual Conference on Manual Control.* Wright-Patterson Air Force Base, OH.

Wherry, R. J., Jr. (1969). The development of sophisticated models of man-machine systems. In *Proceedings of the Symposium on Applied Models of Man-Machine Systems Performance.* Columbus, OH: North American Aviation.

Wherry, R. J., Jr. (1986). *Theoretical development for identifying underlying internal processes. Volume 1. The theory of underlying internal processes.* NAMRL/NADC Joint Report. NADC Report 86105-60. Warminster, PA: Navel Air Development Center.

Whitaker, L. A. (1979). Dual-task interference as a function of cognitive processing load. *Acta Psychologica, 43,* 71-84.

White, S. A., MacKinnon, D. P., & Lyman, J. (1985). Modified Petri net model sensitivity to workload manipulations. *Proceedings of the Twenty-first Annual NASA-University Conference on Manual Control* (pp. 3.1-3.17). Columbus, OH: Ohio State University.

Wickens, C. D. (1976). The effects of divided attention on information processing in manual tracking. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 1-12.

Wickens, C. D. (1980). The structure of attentional resources In R. Nickerson (Ed.), *Attention and Performance VIII* (pp. 239-257). Hillsdale, N.J: Lawrence Erlbaum Associates.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman, & D. R. Davies (Eds.), *Varieties of attention* (pp. 63-102). New York: Academic Press.

Wickens, C. D., & Gopher, D. (1977). Control theory measures of tracking as indices of attention allocation strategies. *Human Factors, 19,* 349-365.

Wickens, C. D., & Kessel, C. (1980). Processing resource demands of failure detection in dynamic systems. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 564-577.

Wickens, C. D., & Yeh, Y-Y. (1985). POCs and performance decrements: A reply to Kantowitz and Weldon. *Human Factors, 27,* 549-554.

Wickens, C. D., Hyman, F., Dellinger, J., Taylor, H., & Meador, M. (1986). The Sternberg memory search task as an index of pilot workload. *Ergonomics, 29,* 1371-1383.

Wickens, C. D., Kramer, A., Vanasse, L., & Donchin, E. (1983). The performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information processing resource. *Science, 221,* 1080-1082.

Wickens, C. D., Mountford, S. J., & Schreiner, W. (1981). Multiple resources, task-hemispheric integrity, and individual differences in time-sharing. *Human Factors, 23*, 211-229.

Wickens, C. D., Zenyuh, J., Culp, V., & Marshak, W. (1985). The effects of voice and manual control mode on dual task performance. In *Proceedings of the 21st Annual Conference on Manual Control* (p. 11.1). Columbus, OH: Ohio State University.

Wierwille, W. W. (1979). Physiological measures of aircrew mental workload. *Human Factors, 21*, 575-593.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement application. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 129-133). Santa Monica, CA: Human Factors Society.

Wierwille, W. W., & Conner, S. A. (1983). Evaluation of twenty workload assessment measures using a psychomotor task in a motion-base aircraft simulation. *Human Factors, 25*, 1-16.

Wierwille, W. W., & Gutmann, J. C. (1978). Comparison of primary and secondary task measures as a function of simulated vehicle dynamics and driving conditions. *Human Factors, 20*, 233-244.

Wierwille, W. W., & Williges, B. H. (1980). *An annotated bibliography on operator mental workload assessment.* (Naval Air Test Center Report No. SY-27R-80). Patuxent River, MD: Naval Air Test Center, System Engineering Test Directorate. (ADA 083636).

Wierwille, W. W., & Williges, R. C. (1978). *Survey and analysis of operator workload* (S-78-101). Blacksburg VA: Systemetrics, Inc.

Wierwille, W. W., & Williges, R. C. (1978). *Survey and analysis of operator workload* (S-78-101). Blacksburg VA: Systemetrics, Inc.

Wierwille, W. W., Casali, J. G., Connor, S. A., & Rahimi, M. (1985). Evaluation of the sensitivity and intrusion of mental workload estimation techniques. In W. Roner (Ed.), *Advances in Man-Machine Systems Research* Volume 2. (pp. 51-127). Greenwich, CT: J.A.I. Press.

Wierwille, W. W., Gutmann, J.C., Hicks, T. G., & Muto, W. H. (1977). Secondary task measurement of workload as a function of simulated vehicle dynamics and driving conditions. *Human Factors, 19*, 557-565.

Wierwille, W. W., Skipper, J., & Reiger, C. (1984). Decision tree rating scales for workload estimation. Theme and variations (NASA-CP-2341). In *Proceedings of the 20th Annual Conference on Manual Control* (pp. 73-84). Washington, DC: NASA.

Williams, L. J. (1982). Cognitive load and the functional field of view. *Human Factors, 24*, 683-692.

Williges, R. C., & Wierwille, W. W. (1979). Behavioral measures of aircrew mental workload. *Human Factors, 21*, 549-574.

Wilson G. F., & O'Donnell, R. D. (1986). Steady state evoked responses: Correlations with human cognition. *Psychophysiology, 23*, 57-61.

Wilson, G.F., O'Donnell, R. D., & Wilson, L. (1983). *Neurophysiological measures of A-10 workload during simulated low altitude missions* (AFAMRL-TR-83-0003). Wright-Patterson AFB, OH: Aerospace Medical Research Laboratory.

Wolf, J. D. (1978). *Crew workload assessment: Development of a measure of operator workload.* (AFFDL-TR-78-165). Wright Patterson AFB, OH: Air Flight Dynamic Laboratory.

Wright, P., Holloway, C. M., & Aldrich, A. R. (1974). Attending to visual or auditory verbal information while performing other concurrent tasks. *Quarterly Journal of Experimental Psychology, 26*, 454-463.

Yeh, Y-Y, & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30*, 111-120.

Yeh, Y-Y., & Wickens, C. D. (1984). *The dissociation of subjective measures of mental workload and performance* (Technical Report EPL-84-2/NASA-84-2). University of Illinois at Urbana-Champaign: Engineering Psychology Research Laboratory.

Yoakum, C. S., & Yerkes, R. M. (Eds.) (1920). *Army mental tests.* New York: Henry Holt and Company.

Zachary, W. (1981). *Cost-benefit assessment of candidate decision aids for Naval Air ASW.* (Analytics Tech Report 1366-C). Willow Grove, PA: Analytics.

Zachary, W., Zaklad, A., & Davis, D. (1987). A cognitive approach to multisensor correlation in an advanced tactical environment. Paper presented at 1st Tri-Service Data Fusion Symposium, Johns Hopkins University, Columbia, MD.

Zaklad, A., Deimler, J., Iavecchia, H. P., & Stokes, J. (1982). *Multisensor correlation and TACCO workload in representative ASW and ASUW environment.* (Analytics Tech Report 1753A). Willow Grove, PA: Analytics.

Zeitlin, L. R., & Finkelman, J. M. (1975). Research note: Subsidiary task techniques of digit generation and digit recall indirect measures of operator loading. *Human Factors, 17*, 218-220.

APPENDIX A. LITERATURE REVIEW OF SECONDARY TASKS

The approach taken to review the vast secondary task literature was to identify any relationships that may exist between secondary task characteristics and primary task characteristics. That is, we classified the types of secondary tasks and primary tasks that have been reported in the literature. We then examined the results of such studies based on the various secondary and primary task configurations. Our reasoning behind this effort was to address a pragmatic question that other researchers have recognized as being very important but overlooked (Chiles & Alluisi, 1979). That is, are operators capable of performing two different tasks concurrently? To illustrate, if either the primary or secondary task exhibits a decrement in performance when they are performed jointly, this finding, at the very least, suggests that operators find it difficult to perform such a dual task configuration. Even though such findings may violate methodological assumptions needed to draw inferences about spare capacity, it provides valuable information about the ability of operators to exhibit time-sharing ability between two tasks. (See Gopher and Donchin [1986] for an overview of the literature that specifically addresses time-sharing ability.) To our knowledge, there have been no published reports that have followed such a scheme for the secondary task literature. Ogden et al. (1979) have provided a basis for such an analysis, but they did not actually complete the analysis.

We recognize that even though the results from this analysis may suggest that certain dual task configurations result in primary task performance decrements, it is misleading to suggest this will be the case in every situation. Rather, it is our intention to alert human factors practitioners to consider the implications of this analysis with respect to their particular situation. This is important as technological advances have increased the complexity of systems such that operators are routinely required to perform more than one task at any time. The secondary task paradigm is a controlled analog of this situation.

We were also interested in identifying any trends from this analysis that suggested performance changes that are sensitive to secondary task characteristics as a function of primary task characteristics.

Identification of Secondary Task Literature

The primary reference sources used to identify relevant secondary task articles published prior to 1980 were Ogden et al. (1979) and Wierwille and Williges (1980). For relevant articles published after 1979, we identified key people in this area through our OWL Information System database and sent requests to such individuals for their most recent articles on operator workload. We also searched relevant journals (e.g., *Human Factors*) and proceedings of conferences and meetings (e.g., *NASA/University Conferences on Manual Control*) for recent studies. As a result of this effort, we were able to obtain 147 studies for review. Of these, seven were excluded from our analysis because they lacked sufficient information to interpret their results. Four studies were also excluded because they dealt with multiple task batteries in which no attempt was made to examine dual task performance. We were left with 136 articles and 181 experiments to be analyzed. This literature base is a comprehensivoly representative sample of the secondary task literature.

Classification Scheme for Primary and Secondary Tasks

Classification of secondary and primary tasks characteristics was attempted following the major classes reported in the literature (O'Donnell & Eggemeier, 1986; Ogden et al, 1979). However, due to the variety of secondary and primary tasks that have been employed in studies, it was necessary to expand previous classification schemes as well as to identify particular tasks that have received extensive use (e.g., the Stemberg memory task). Listed below is the classification scheme we developed with descriptions for each category. This list represents the entire range of 26 tasks that we were able to identify for secondary and primary task characteristics based on our review.

- **Choice Reaction Time Task** – the subject is presented with more than one stimulus and must generate a different response for each one. Visual or auditory stimuli may be employed and the response mode is usually manual. It is theorized that choice reaction time imposes both central processing and response selection demands.
- **Simple Reaction Time Task** – the subject is presented with one discrete stimulus (either visual or auditory) and generates one response to this stimulus, minimizing central processing and response selection demands.
- **Driving Task** – the subject operates a driving simulator or actual motor vehicle. Such a task involves complex psychomotor skills.
- **Randomization Task** – the subject must generate a random sequence of numbers, for example. It is postulated that with increased workload levels subjects will generate repetitive responses (i.e., lack randomness in responses).
- **Tracking Task** – the subject must follow or track a visual stimulus (target) which is either stationary or moving by means of positioning an error cursor on the stimulus using a continuous manual response device. Central-processing and motor demands vary depending on the order of control dynamics for the device used by the subject to control the error cursor.
- **Monitoring Task** -- the subject is required to maintain attention to a visual display and to detect the occurrence of a stimulus (signal) from among several alternatives (neutral events). The task is not intermittent but continuous. Monitoring tasks are generally assumed to impose a heavy load on perceptual processes.
- **Time Estimation Task** -- the subject keeps track of time either by generating a specific time interval or by estimating the duration of a time interval at its conclusion. Typically, subjects are required to generate 10 second time intervals (time production procedure) and it is assumed under high workload conditions that subjects will underestimate the passage of time as reflected by their responses (i.e., longer time estimates).
- **Memory Task** -- there are a variety of memory tasks which employ a number of different types of materials and specific requirements. For example, the subject is required to recall in any order a list of words previously memorized (free recall paradigm) or is required to recognize previously memorize words from a list of words (recognition recall paradigm). These tasks are typically assumed to impose heavy demands on central-processing resources.

- **Mental Mathematics Task** – the subject must perform mental arithmetic operations such as addition, subtraction, and multiplication. These tasks are generally considered to place heavy demands on central-processing resources.
- **Michon Interval Production Task** – the Michon paradigm of interval production requires the subject to generate a series of regular time intervals by executing a motor response (i.e., a single finger tap every 2 sec.). No sensory input is required. This task is thought to impose heavy demand on motor output/response resources. It has been demonstrated with high demand primary tasks that subjects exhibit irregular or variable tapping rates.
- **Sternberg Memory Task** – the Sternberg memory task is a commonly used memory task. The subject is presented with a set of digits or letters to memorize. Subsequently, the subject is presented with a test digit or letter and must judge whether this digit was contained in the previous memorized set. It is theorized that the Sternberg memory task aids in workload assessment by distinguishing between primary task central processing effects from primary task stimulus encoding/response execution effects.
- **Lexical Decision Task** – typically, the subject is briefly presented with a sequence of letters and must judge whether this letter sequence forms a word or a non-word. This task is thought to impose heavy demands on semantic memory processes.
- **Distraction Task** – the subject performs a task which is executed in a fairly automatic way such as counting aloud. Such a task is intended to distract the subject in order to prevent the rehearsal of information that may be needed for the primary task.
- **Problem Solving Task** – the subject engages in a task which requires verbal or spatial reasoning. For example, the subject might attempt to solve anagram or logic problems. This class of tasks is thought to impose heavy demands on central processing resources.
- **Identification/Shadowing Task** – The subject identifies changing symbols (digits and/or letters) that appear on a visual display by writing or verbalizing, or repeating a spoken passage as it occurs. Such tasks are thought to impose demands on perceptual processes (i.e., attention).
- **Detection Task** – the subject must detect a specific stimulus or event which may or may not be presented with alternative events. For example, to detect which of 4 lights is flickering. The subject is usually alerted by a warning signal (e.g., tone) before the occurrence of such events, therefore attention is required intermittently. Such tasks are thought to impose demands on perceptual processes.
- **Classification Task** – the subject must judge whether symbol pairs are identical in form. For example, to match letters either on a physical level (AA) or on a name level (Aa). Depending upon the requirements of the matching task, the task can impose demands on perceptual processes (physical match) and/or cognitive processes (name match or category match).
- **Psychomotor Task** – the subject must perform a psychomotor task such as sorting different types of metal screws by size. Tasks of this nature are thought to reflect psychomotor skills.
- **Spatial Transformation Task** – the subject must judge whether information (data) provided by an instrument panel or radar screen - matches information which is spatially depicted by pictures or drawings of aircraft. This task involves perceptual and cognitive processes.

- **Speed Maintenance Task** -- the subject must operate a control knob to maintain a designated constant speed. This task is a psychomotor type task.

- **Production/Handwriting Task** -- the subject is required to produce spontaneous handwritten passages of prose. With primary tasks that impose a high workload, subject's handwriting is thought to deteriorate (i.e., semantic and grammatical errors) under such conditions.

- **Card Sorting Task** -- the subject must sort playing cards by number, color, and/or suite. Depending upon the requirements of the card sorting rule, the task can impose demands on perceptual and cognitive processes.

- **Three Phase Code Transformation Task** -- the subject operates the 3P-Cotran which is a workstation consisting of three indicator lights, a response board for subject responses and a memory unit that the subject uses to save his/her responses. The subject must engage in a 3 phase problem solving task by utilizing information provided by the indicator lights and recording solutions onto the memory unit. it is a synthetic work battery used to study work behavior and sustained attention.

- **Multi-Task Performance Battery (MTPB)** -- the subject operates a workstation consisting of display panels and response control panels for six different tasks (choice RT, monitoring, mental math, identification, problem solving, and tracking). The task battery is designed to involve perceptual, cognitive, stimulus encoding, and response selection processes.

- **Occlusion Task** -- the subject's view of a visual display is obstructed (usually by a visor). These obstructions are either initiated by the subject or imposed by the experimenter in order to determine the viewing time needed to perform a task adequately.

- **Simulated Flight** -- the flight simulators used in the studies that were part of our analysis were typically commercially available training simulators (e.g., Singer-Link GAT-1B). Depending on the purpose of the particular study, the subject was required to perform various maneuvers (e.g., landing approaches) under different types of conditions such as instrument flight rules or simulated crosswind conditions.

Measures Used with Primary and Secondary Tasks

The complexity of the results found with different secondary and primary task pairings can be partly attributed to the different and numerous performance measures that have been used with these tasks. Also, studies which have used the same primary and secondary task pairings have either used different measures of performance or reported different results from the same task measures. In Table A-1, we have listed several of the frequently reported measures that have been recorded with primary and secondary tasks. These measures are organized according to the task classification just presented.

Table A-1. Measures utilized to quantify performance on primary and secondary tasks.

TASK	MEASURE
Choice Reaction Time Task	Mean (median) RT for correct responses Mean (median) RT for incorrect responses Number (%) correct responses Number (%) incorrect responses
Simple Reaction Time Task	Mean (median) RT for correct responses Number (%) correct responses
Driving Task	Total time to complete a trial Number of acceleration rate changes Number of gear changes Number of footbrake operations Number of steering reversals Number of obstacles hit High pass steering (standard) deviation Yaw (standard) deviation Lateral (standard) deviation
Randomization Task	% redundancy score (bits of information)
Tracking Tasks	Integrated errors in volts(root mean square error) Total time on target Total time of target Number of times of target Number of target hits
Monitoring Tasks	Number (%) of correct detections Number (%) of incorrect detections Number (%) of errors of omission Mean (median) RT for correct detections Mean (median) RT for incorrect detections
Memory Tasks	Mean (median) RT for correct responses Number (%) of correct responses Number (%) errors of omission Number (%) of incorrect responses
Mental Mathematics Tasks	Number (%) of correct responses Mean (median) RT for correct responses Number (%) of incorrect responses
Michon Interval Tapping Task	Mean interval per trial Standard deviation of interval per trial Sum of differences between successive intervals per minute of total time
Sternberg Memory Task	Slopes and intercepts for RT data (See memory tasks)
Lexical Decision Task	Mean RT for correct responses

Table A-1. Measures utilized to quantify performance on primary and secondary tasks (Cont.).

TASK	MEASURE
Problem Solving Tasks	Number (%) of correct responses Number (%) of incorrect responses Mean (Median) RT for correct responses
Identification/Shadowing Task	Number of words correct/minute Number of digits spoken Mean time interval between spoken digits Number of errors of omission
Detection Tasks	Mean RT for correct detections Number (%) of correct detections
Classification Tasks	Mean (median) RT for physical match Mean (median) RT for category match Number (%) errors for physical match Number (%) errors for category match
Psychomotor Tasks	Number of completed items
Spatial Transformation Tasks	Mean RT for correct responses Number (%) of correct responses Number (%) of incorrect responses
Occlusion Tasks	Mean voluntary occlusion time Percent looking time/total time
Spontaneous Writing	Number of semantic and grammatical errors
Card Sorting Tasks	Number of cards sorted Number (%) of incorrect responses
3P-Cotran Task	Mean (median) RT for different phases of response required Number of errors (resets) for different phases of response required
MTPB	Mean (median) RT for correct detections Number (%) of correct detections Number of problems attempted
Simulated Flight	Number of vertical accelerations Mean error from required altitude Root mean square localizer error Root mean square glide-slope error Number of control movements Pitch high-pass mean square Roll high-pass mean square

Analysis Scheme for Primary and Secondary Tasks

In order to analyze the results from the studies reviewed, we established conventions to provide a framework for tabulating such complex findings. We worked with the premise that it was important to report any indication that dual task pairings resulted in a performance decrement on one or both of the tasks (primary or secondary). Based on this premise, we formulated the following criteria for priorities on the results reported in each study:

- Measures which revealed differences between dual-task and single-task performance were tabulated and preferred over results for measures which showed performance stability for the same task.
- Measures which revealed decrements for dual-task versus single-task performance were tabulated and preferred over results for measures which showed performance enhancement for the same task.
- Measures which revealed dual-task performance decrements with experimental manipulations (i.e., different sound levels of noise, different levels of task demand for either secondary or primary tasks, etc.) were tabulated and preferred over results for measures which showed no effects for the same task.

Analysis of Secondary Task Literature

A systematic approach was undertaken to characterize the wealth of information contained in the experiments reviewed. The approach involved several steps.

We first characterized the studies reviewed according to the primary and secondary tasks employed. Table A-2 and Table A-3 contain the results of this effort. As seen in Table A-2, it is evident that the majority of experiments have involved a select number of secondary tasks. The first four secondary tasks listed in Table A-2 represent over 50% of the total secondary tasks that comprised our sample. Similarly with respect to primary tasks, the first three tasks listed in Table A-3 represent over 50% of the total primary tasks in our sample. It is also evident in examining Table A-3 that a small percentage of the studies reviewed have employed primary tasks which can be characterized as realistic. That is, primary tasks typically do not involve multiple sencory input and several types of operator actions and responses (i.e., driving and simulated flight). Such findings reflect the academic interests of researchers who have utilized the secondary task paradigm.

We further characterized the articles according to the particular primary-secondary configuration employed. This was accomplished in two complementary ways. We examined secondary tasks with respect to the various primary tasks that have been used in association with each secondary task. Similarly, we examined primary tasks with respect to the various secondary tasks that have been used in association with each primary task. We were interested in identifying any trends in the results across similar dual task pairing experiments that would suggest particular dual task pairings are not advantageous for the operator (i.e., performance decrements for both secondary and primary tasks). Attachment 1

Table A-2. Number of experiments to utilize secondary tasks.

Secondary Task	Number of Experiments
Monitoring Tasks	41
Memory Tasks	32
Choice Reaction Time Tasks	25
Mental Mathematics Tasks	18
Tracking Tasks	12
Simple Reaction Time Tasks	11
Michon Interval Production Task	11
Identification Tasks	9
Problem Solving Tasks	8
Time Estimation Tasks	7
Detection Tasks	6
Sternberg Memory Task	5
Randomization Task	5
Occlusion Tasks	4
Psychomotor Tasks	3
Card Sorting Tasks	2
Spontaneous Handwriting Task	1
Aircraft Navigation Task	1
Spatial Transformation Tasks	1
MTPB (monitoring tasks)	1
Distraction Tasks	1
TOTAL	203

The total 203 represents the instances that these secondary tasks were used in experiments. Several studies used more than one secondary task in a single experiment. The total 203 is based on 180 experiments.

Table A-3. Number of experiments that utilized primary tasks with secondary tasks.

Primary Task	Number of Experiments
Tracking Tasks	48
Memory Tasks	26
Monitoring Tasks	23
Choice Reaction Time Tasks	18
Driving Tasks	17
Simulated Flight	8
Detection Tasks	7
Problem Solving Tasks	7
Identification Tasks	5
Classification Tasks	4
Mental Mathematics Tasks	4
Simple Reaction Time Tasks	3
Psychomotor Tapping Tasks	3
Card Sorting Tasks	1
Spatial Transformation Task	1
Sternberg Memory Task	1
Distraction Tasks	1
Lexical Decision Task	1
3P Contran Task	1
MTPB	1
Psychomotor Tasks	1
TOTAL	181

The total 181 represents the experiments reported in the 147 articles that we reviewed for our analysis.

contains the results from the analysis of individual secondary tasks each paired with various primary tasks. Attachment 2 contains the results from the complementary analysis of individual primary tasks each paired with various secondary tasks.

Discussion of Secondary Task Analysis Results

Perusal of A-4 reveals the complex nature of the results that have been reported with various primary-secondary task pairings. With respect to practical considerations, the results depicted in Table A-4 reflect the need on the part of human factors practitioners to examine the specific demands placed on operators whenever there are system requirements to perform several tasks at once. This point is illustrated by examining the major classes of secondary tasks that are depicted in Table A-4.

Inspection of monitoring secondary task experiments reveal several interesting trends. With monitoring-monitoring dual task pairings, performance on the primary task seems to decline consistently across experiments with one exception. A somewhat similar finding is shown in these same experiments with respect to the monitoring tasks designated as secondary tasks. As these experiments did not place a greater emphasis on either primary or secondary monitoring tasks with respect to maintaining performance levels such findings are possibly due to the high task demands that two monitoring tasks combined placed on perceptual processes. When one examines the tracking-monitoring dual task pairing results, a somewhat different picture emerges. The primary tracking task results exhibit across experiments an almost equal split between stable performance and degraded performance. However, the experiments that reported degraded performance for the primary tracking task all placed equal emphasis on both primary and secondary task performance. This may have contributed to subjects' poor performance on the tracking tasks because subjects may have formed inappropriate strategies for handling the dual task pairing. While those experiments that reported primary tracking task performance as stable, a greater emphasis was placed on tracking performance for three out of seven experiments listed in this category. With respect to the monitoring secondary task results in this dual task configuration,performance appears stable when the monitoring task is auditory in nature. But when the monitoring task is visual, six out of seven experiments reported a performance decrement on the monitoring portion. Such findings seem to indicate that visual tracking-visual monitoring dual task pairing can lead to performance decrements. This is the case especially on the visual monitoring portion probably because of the combined visual load placed on subjects by the two tasks.

For secondary memory task experiments, the results seem to exhibit a trend across experiments that indicates prior experience with a task is an important factor for dual task performance. With driving-memory dual task situations, only the memory portion revealed a performance decrement. As these experiments involved experienced drivers with greater emphasis placed on driving performance, these factors probably contributed for such driving stability but at the expense of the memory task. In contrast, tracking-memory dual task situations resulted in both tasks exhibiting poor performance for most experiments. These experiments typically involved college students and their inability to perform this dual task configuration reflects lack of experience, even though researchers try to provide for task mastery.

With respect to choice reaction time secondary task experiments, it is evident that the dual task pairing consisting of tracking-choice reaction time results in poor performance on both tasks with one exception. The complexity of this dual task situation (i.e., task demands on central processing, response selection and motor responses for subjects) probably contributes to such poor overall performance. Similar results are found with experiments that employed mental mathematics secondary tasks. As seen in Table A-1, dual task pairings with mental mathematics as the secondary task results in poor performance on both tasks for most experiments. As mental mathematics can be considered a relatively complex set of cognitive operations, its pairing with almost any primary task configuration except simple tasks (e.g., tapping) or highly practiced tasks (e.g., driving) results in poor overall dual performance.

The above descriptions illustrate the complex results found with all secondary task studies. The results reflect the complex interactions between the salient factors that influence performance in dual task situations.

Conclusions

The complexity of the results just described may seem, at first, to be beyond simple conclusions or implications. However, several important issues can be derived concerning the use of secondary tasks as an OWL technique and dual task performance in general.

With respect to secondary tasks as a workload estimation technique, the results described show that secondary tasks can interfere with primary task performance. As a result, inferences concerning spare capacity with a primary task become difficult to interpret. A solution to this problem is to employ tasks as secondary tasks that are inherently part of a multitask system. Under these circumstances, a wealth of information is gained even though the primary task may show performance changes. Because any change in performance, whether on the designated secondary or primary task of a system, provides valuable insight concerning the operator's capabilities and limits in using the system. It is for this reason that the embedded secondary task technique is offered in Chapter 6 as the technique of choice in system design and development environments.

Another important implication that is derived from our analysis is that secondary tasks can result in changes in primary task performance that seem to be reflective of subjects' inappropriate strategies with respect to the dual task situation. Subjects' performance on the primary task seems to be degraded because the introduction of the secondary task has changed the nature of the situation with respect to primary task demands. If your interests are to quantify the spare capacity with respect to a primary task, then such changes are clearly troublesome. To prevent these possible circumstances, it is necessary to use secondary task techniques that do not intrude on primary task performance. Several secondary task techniques are offered Chapter 6 that minimize this potential confounding. They have been demonstrated in some applied settings not to intrude on operators' performance with complex systems and to be sensitive to workload levels on such systems.

For dual task performance, the secondary task literature provides, though in some cases unintentionally, valuable information on the critical factors that hinder multi-task performance. These factors are:

- inappropriate operator strategies with respect to meeting the task demands of several tasks at once,
- the potency of certain types of tasks (e.g., mental mathematics) to hinder the ability of operators to perform any additional tasks that may be required, and
- the combined task demand effects of certain task configurations (e.g., monitoring-monitoring) are such that they overload the operator when performed together.

The human factors practitioner needs to be aware of these factors in order to ensure that performance on complex systems does not suffer from such factors.

APPENDIX A – ATTACHMENT 1

Secondary Task Experiments with Respect to Secondary Task Characteristics

Attachment 1 is shown on the following pages and contains the results from the analysis of individual primary tasks with respect to secondary task pairings. The table is organized in the following manner:

- The particular primary task examined is identified in the far left-hand column header of each page. It is indicated with the letter "P" preceding the primary task characteristic, for example P-monitor.
- The secondary task pairings associated with the particular primary task are listed below the primary task header in the far left-hand column.
- The experiments that employed these particular dual task pairings are listed by first author and year for cited article reference. They appear in the table under the appropriate column with respect to the results for the primary task as well as secondary task. If an experiment only reports the results for eit the primary or secondary task then the experiment is listed only once under the appropriate column for the results reported.
- Based on the conventions/rules described in Appendix B, experiments are listed under the appropriate column headers as follows:

  P- signifies primary task measures were stable in dual task pairings

  P Down signifies primary task measure(s) exhibited a decrement in dual task pairings

  P Up signifies primary task measure(s) exhibited an increment in dual task pairings

  S- signifies secondary task measures were stable in dual task pairings

  S Down signifies secondary task measure(s) exhibited a decrement in dual task pairings

  S Up signifies secondary task measure(s) exhibited an increment in dual task pairings

- Each experiment is listed in the following sequential manner:

  First author's last name only for the cited reference article.

  For example, "Domic"

  If one author then the author's last name is underlined.

  For example, "<u>Domic</u>"

  The year the cited reference article was published.

  For example, Domic80

  If the cited reference article contained more than one experiment then the particular experiment is indicated within parenthesis.

  For example, Domic80(1)

The primary task's mode for stimulus information:

V=visual input

A=auditory input

T=cutaneous input

A+V=both auditory and visual simultaneously

A/V=both auditory and visual but not simultaneously

-- not appropriate

For example, Domic80(1)v

The secondary task's mode for stimulus information:

V=visual input

A=auditory input

T=cutaneous input

A+V=both auditory and visual simultaneously

A/V=both auditory and visual but not simultaneously

-=not appropriate

For example, Domic80(1)va

The emphasis placed on maintaining performance for either primary or secondary tasks during dual task pairings as specifically stated in the article or implied by payoff matrices (e.g., $10 for high scores on the primary task).

P=primary task emphasized

S=secondary task emphasized

Blank=both secondary and primary are emphasized or the authors do not specifically state the performance emphasis placed on subjects therefore assumed equal emphasis for both primary and secondary tasks

For example, Domic80(1)vap

If the experiment contained data that allowed the determination that either the secondary or primary task performance changed (i.e., increment or decrement) or was stable in dual task pairings but was not specifically addressed by the authors, this is indicated under the appropriate primary or secondary result column header as interpolated.

For example, Domic80(1)vapintrp

S-MICHON	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MONITOR		Shingledecker 83v-		Shingledecker 83v-		
PROBLEM SOLVE		Michon64(2) v- (multiple task comparisons)			Michon64(2) v- (multiple task comparisons)	
SIMPLE RT			Vroon73avp		Vroon73avp	
DETECTION		Michon64(2) v- (multiple task comparisons)			Michon64(2) v- (multiple task comparisons)	
PSYCHO-MOTOR		Michon64(2) v- (multiple task comparisons)			Michon64(2) V- (multiple task comparisons)	
STERNBERG		Shingledecker 83v-		Shingledecker 83v-		
FLIGHT SIMULATE	Wierwille85 (2,3,4)v-			Wierwille85 (3,4)v-	Wierwille85 (2)v-	
DRIVING	Brown67v-p				Brown67v	
TRACKING		Shingle-decker 83v-			Shingle-decker 83v-	
CHOICE RT		Michon64 (1) v?-			Michon64(1) v?- Michon66vv	
MEMORY		Roediger75 (1,2)vv		Roediger75 (1,2)vv		
MENTAL MATH		Michon64(2) v- (multiple task comparisons)			Michon64(2) v- (multiple task comparisons)	

S-CARD SORTING	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MEMORY		Murdock65 (1)av Murdock65 (2)avs			Murdock65 (1)av Murdock65 (2)avp	

S- DETECTION	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Wickens81vv			Wickens81vv	
MEMORY		Hoffman83 vvp			Shulman71 (2)av Shulman71 av Hoffman83 vvp	
MONITOR		Tichner72vv				
DETECTION		Wickens81 vv			Wickens81 vv	
CLASSIFY	Williams82 vvp				Williams82 vvp	

S-CHOICE RT	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Girouard84 (1,2)va Klapp84ap- Trumbo72 (1,2)va Benson65vvp Wempe68va Israel80vap	Gopher77vv	Gopher77vv	Girouard84 (1,2)va Klapp84ap- Benson65vvp Israel80vap Damos78vvp	
CHOICE RT	Becker76va Ells73va	Schvaneveldt 69(1,2)vv			Becker76va Ells73va Schvaneveldt 69(1,2)vv	
MEMORY		Logan70 (1,2,3)avp			Logan70 (1,2,3)avp	
MONITOR		Smith69(1,2) av			Smith 69 (1,2)av Krol71vap	
PROBLEM SOLVE	Fischer75 avp intrp				Fischer75 avp intrp	
FLIGHT SIMULATIO N	Bortolussi87 vvp Bortolussi86 vv				Bortolussi87 vvp Bortolussi86 vv	
DRIVING		Brown69va Allen75vv		Allen75vv	Brown69va	
LEXICAL DECISION		Becker76vap			Becker76vap	

S- DISTRACT	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MEMORY		Glanzer66 a+v,v				

S-IDENTIFY	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Gabay77vv		Gabay77vv		
MEMORY		Mitsuda68aa				
MONITOR					Savage78vv p	
DETECTION		Price75vvp			Crice75vvp	
IDENTIFY	Allport72avp			Allport72avp		
DRIVING		Hicks79vvp Wierwille78 vvp			Wierwille77 vvp Wierwille75 vvp	
SPATIAL TRANSFOR MATION		Fournier76 vap			Fournier76 vap	

S-MENTAL MATH	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING	Huddleston71 vvp	McLeod73va Wickens81 Bahrick54va Chiles70 Heimstra70 va		Bahrick54va Heimstra70 va	McLeod73va Wickens81 Huddleston71 vvp	Chiles70
CHOICE RT		Fisher75vap Keele67va Chiles70vv			Fisher75vap Keele67va Schouten62 av?p	
MEMORY		Silverstein71 vv Roediger75 (3)vv				
MONITOR		Kahneman67 vap Chiles70vv			Kahneman67 vap Chiles70vv	
SIMPLE RT		Chiles70vv				
DETECTION		Jaschinski82 va			Jaschinski82 va	
DRIVING	Wetherall81 va Brown61vap				Wetherall81 va Brown61vap	
TAPPING	Kantowitz76 (1,2)vap				Kantowitz76 (1,2)vap Kantowitz74 vvp	

S-MEMORY	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING	Finkelman54 va Zeitlin75va	Trumbo71 (1,2)vv Noble67va Heimstra70 va Wickens80va 'Huddleston 71vvp Wickens81va	Tsang84 v,v/ap	Trumbo71 (1,2)vv Noble67va	Wickens80 va Huddleston 71vvp Wickens81 va Zeitlin75va Finkelman54 va,Heimstra 70va,Tsang 84v,v/ap	
CHOICE RT		Broadbent65 ta Keele73vv			Broadbent65 ta	
MEMORY		Broadbent62 va Chow75 (1,4)vv (3)va			Shulman71 (1)av	
MENTAL MATH	Mandler73 avp				Mandler73 avp	
MONITOR	Moskowitz74 aa Chechile79	Chiles79 v+a,v			Chiles79 v+a,v Moskowitz74 aa intrp Chechile79v+ a,v	
PROBLEM SOLVE	Daniel65va	Stager72vv				
DETECTION		Wickens81va			Wickens81va	
IDENTIFY		Klein76va			Allport72 a,v/ap	
CLASSIFY		Wickens81va			Wickens81va	
DISTRACT		Broadbent62 va				
DRIVING	Brown62,66 and 68 vap 62 intrp Wetherall81 va Brown61vap			Brown65 vap	Brown62,66 and 68 vap 62intrp Wetherall81 va Brown61vap	

S-MTPB	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
3P-COTRAN		Alluisi71			Alluisi71	

S-MONITOR	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING	Schori73vv Bell78vap, Gabriel68vv Kelly67vv Huddleston 71vvp Figarola66va Kyriakides77 vvp	Bergeron68 vv Herman65aa Putz74vv Kramer84 v,v/a Malmstrom 83va Heimstra70 vv Monty65vv		Figarola66va Kramer84 v,v/a Malmstrom 83va	Bergeron68 vv,Bell78vap Herman65aa Monty65vv Kyriakides 77vvp,Putz 74vv.Gabriel 68vv,Schori 73vv.Kelly 67vv, Huddiaston71vvp	Heimstra70 vv
CHOICE RT	Boggs68vap	Hilgendorf 67vap			Hilgendorf 67vap Boggs68vap	
MEMORY	Tyler74va	Mitsuda68aa Lindsay69aa Chow75(2)va		Lindsay69aa		
MENTAL MATH	Domic80 v?vp				Domic80 v?vp Chiles79vv	
MONITOR		Long76(1,2) va Fleishman65 vv Hohmuth70 (1)av/va Goldstein78 vv Chechile79 v+a,v Stager71vas	McGrath65 a/v,a+v	Stager71 vas	Long76(1,2) va Chechile79 v+a,v Hohmuth70 (1,3)av/va	
PROBLEM SOLVE	Wright74 (1)vvp				Wright74(1) vvp Chiles79vv	
DETECTION		Dewar76vv Tyler74va			Tyler74va	
IDENTIFY	Domic80 (1)vap (2)vvp				Domic80 (1)vap (2)vvp Chiles79vv	
FLIGHT SIMULATION				Soliday65vv		
DRIVING	Brown62vap int p Wetherell81 va Hoffman66 vap Brown67vvp	Brown65vap		Hoffman66 vap	Brown62vap intrp Brown65vap	Brown67 vvp

S - PROBLEM SOLVING	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
PROBLEM SOLVING					Chiles70	
DRIVING		Wetherall81 va			Wetherall81 va	
TRACKING		Trumbo67va Figarola66vv		Trumbo67va Figarola66vv		
CHOICE RT					Schouten62 avp	
MEMORY		Trumbo70vv			Trumbo70vv	
MONITOR	Gould67vvp Smith66vv					Smith66vv

S - NAVIGATE	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
FLIGHT SIMULATION	Lewis65 vvp (field study)				Lewis65 vvp (field study)	

S - OCCLUSION	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MONITOR	Gould67vvp					
DRIVING	Farber72vvp	Senders67vv Hicks79p			Senders67vv Farber72vvp	

S - RANDOMIZATION	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING	Zeitlin75	Truijens76va		Zeitlin75	Truijens76va	
MEMORY		Trumbo70v-				
CARD SORTING	Baddeley66 v-				Baddeley66 v-	
DRIVING	Wetherall81 v-					

S-TRACKING	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING	Mackenzie72vvp	Hess74vv Wickens80vv Wickens81vv	Tsang84vvp		Wickens81vv Tsang84vvp	Mackenzie72vvp
CHOICE RT		Lauber76vv Whitaker79 vvp			Whitaker79 avp Navon82ava	
MEMORY		Johnston70 av			Johnston70 av	
MONITOR	Griffiths71 av				Griffiths71 av	
PROBLEM SOLVING	Wright74 (2)vvp				Wright74 (2)vvp	
SIMPLE RT		Schmidt84av			Schmidt84av	
DETECTION		Wickens81vv			Wickens81vv	
CLASSIFY		Wickens81vv			Wickens81vv	

S-SIMPLE RT	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Wickens77vv Klapp84va sp Heimstra70 vv Kelly85va			Wickens77vv	Heimstra70 vv
CHOICE RT	Becker76vap				Becker76vap	
MEMORY					Martin74 (1,2)avp	
DETECTION			Laurell78 va intrp		Laurell78 va intrp	
CLASSIFY	Comstock73 vap	Miller75 v,a/vp			Miller75 v,a/vp Comstock73 vap	
DRIVING			Laurell78va intrp		Laurell78va intrp Lisper73va	
LEXICAL DECISION		Becker76vap			Becker76vap	

S-PSYCHO-MOTOR	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Wickens76 Bergeron68			Bergeron68	
CHOICE RT					Schouten62 avp	

S-STERNBERG	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Wickens85 vap	Briggs72 (2)va		Wickens85 vap Briggs72 (1,2)va	
CHOICE RT		Hart85 v/e,v/a				
DRIVING		Wetherell81 va				

S-TIME ESTIMAT.	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MONITOR		Liu87v-			Liu87v-	
FLIGHT SIMULATION	Bortolussi87 v-p Bortolussi86 v- Wierwille85 (1,2,3,4)				Bortolussi87 v-p Bortolussi86 v- Wierwille85 (1,2,3,4)	

S-SPONT. WRITING	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
CHOICE RT					Schouten62 avp	

S-SPATIAL TRANSFORMATION	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING					Vidulich85 v,v/ap	

APPENDIX A. ATTACHMENT 2

Secondary Task Experiments with Respect to Primary Task Characteristics

Attachment 2 is shown on the following pages and contains the results from the analysis of individual primary tasks with respect to secondary task pairings. The table is organized in the following manner:

- The particular primary task examined is identified in the far left-hand column header of each page. It is indicated with the letter "P" preceding the primary task characteristic, for example P-monitor.
- The secondary task pairings associated with the particular primary task are listed below the primary task header in the far left-hand column.
- The experiments that employed these particular dual task pairings are listed by first author and year for cited article reference. They appear in the table under the appropriate column with respect to the results for the primary task as well as secondary task. If an experiment only reports the results for either the primary or secondary task then the experiment is listed only once under the appropriate column for the results reported.
- Based on the conventions/rules described in Appendix B, experiments are listed under the appropriate column headers as follows:

  P- signifies primary task measures were stable in dual task pairings

  P Down signifies primary task measure(s) exhibited a decrement in dual task pairings

  P Up signifies primary task measure(s) exhibited an increment in dual task pairings

  S- signifies secondary task measures were stable in dual task pairings

  S Down signifies secondary task measure(s) exhibited a decrement in dual task pairings

  S Up signifies secondary task measure(s) exhibited an increment in dual task pairings
- Each experiment is listed in the following sequential manner:

  First author's last name only for the cited reference article.

  For example, "Dornic"

  If one author then the author's last name is underlined.

  For example, "Dornic"

  The year the cited reference article was published.

  For example, Dornic80

  If the cited reference article contained more than one experiment then the particular experiment is indicated within parenthesis.

  For example, Dornic80(1)

The primary task's mode for stimulus information:

V=visual input

A=auditory input

T=cutaneous input

A+V=both auditory and visual simultaneously

A/V=both auditory and visual but not simultaneously

-- not appropriate

For example, Domic80(1)v

The secondary task's mode for stimulus information:

V=visual input

A=auditory input

T=cutaneous input

A+V=both auditory and visual simultaneously

A/V=both auditory and visual but not simultaneously

-=not appropriate

For example, Domic80(1)va

The emphasis placed on maintaining performance for either primary or secondary tasks during dual task pairings as specifically stated in the article or implied by payoff matrices (e.g., $10 for high scores on the primary task).

P=primary task emphasized

S=secondary task emphasized

Blank=both secondary and primary are emphasized or the authors do not specifically state the performance emphasis placed on subjects therefore assumed equal emphasis for both primary and secondary tasks

For example, Domic80(1)vap

If the experiment contained data that allowed the determination that either the secondary or primary task performance changed (i.e., increment or decrement) or was stable in dual task pairings but was not specifically addressed by the authors, this is indicated under the appropriate primary or secondary result column header as interpolated.

For example, Domic80(1)vapintrp

P-CHOICE RT	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
PSYCHO-MOTOR					Schouten62 avp	
STERNBERG		Hart85 v/a,v/a				
SPONT. WRITING					Schouten62 avp	
TRACKING		Looper76vv Whitaker79 avp			Whitaker79 avp Hansen82avs	
CHOICE RT	Becker76va Ells73va	Schvaneveldt 69(1,2)vv			Becker76va Ells73va Schvaneveldt 69(1,2)vv	
MEMORY	Broadbent65 ts Keele73vv				Broadbent65 ts	
MENTAL MATH		Fisher75vap Keele67va Chiles70vv			Fishervap Keele67va Schouten62 av?p	
MICHON		Michon64(1) v?-			Michon64(1) v?- Michon66vv	
MONITOR	Boggs68vap	Hilgendorf67 vap			Hilgendorf67 vap Boggs68vap	
FROBLEM SOLVING					Schouten62 avp	
SIMPLE RT	Becker76 vap				Becker76 vap	

P-3PCOTRAN	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MTPB (Dual combinations)		Alluisi71			Alluisi71	

P-CARDSORT	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
RANDOMIZE	Baddeley66 v-				Baddeley66 v-	

P- CLASSIFY	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Wickens81vv			Wickens81vv	
MEMORY		Wickens81va			Wickens81va	
SIMPLE RT	Comstock73 vap	Miller75 v,a/vp			Comstock73 vap Miller75 v,a/vp	
DETECTION	Williams82 vvp				Williams82 vvp	
CLASSIFY		Wickens81vv			Wickens81vv	

P- DETECTION	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Wickens81vv			Wickens81vv	
MEMORY		Wickens81va			Wickens81va	
MENTAL MATH		Jaschinski82 va			Jaschinski82 va	
MICHON		Michon64(2) v- (multiple task comparisons)			Michon64(2) v- (multiple task comparisons)	
MONITOR		Tyler74va Dewar76vv			Tyler74va	
SIMPLE RT			Laurell78va Extrap		Laurell78va Extrap	
DETECTION		Wickens81vv			Wickens81vv	
IDENTIFY		Price75vvp			Price75vvp	

P- DISTRACT	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MEMORY		Broadbent62 va				

P-DRIVING	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
IDENTIFY		Hicks79vvp Wierwille78 vvp			Wierwille77 vvp Wierwille75 vvp	
STERNBERG		Wetherall81 va				
OCCLUSION	Farber72vvp	Sanders67vv Hicks79p			Farber72vvp Sanders67vv	
CHOICE RT		Allen75vv Brown69va		Allen75vv	Brown69va	
MEMORY	Brown68vap Brown66vap Brown62vap intrp Brown61vap Wetherall61 va			Brown65vap	Brown68vap Brown66vap Brown62vap intrp Brown61vap Wetherall81 va	
MENTAL MATH	Wetherall81 va Brown61vap				Wetherall81 va Brown61vap	
MICHON	Brown67v-p				Brown67v-p	
MONITOR	Brown67vvp Brown62vap intrp Wetherall81 va Hoffman66 vap	Brown65vap		Hoffman66 vap	Brown62vap intrp Brown65vap	Brown67vvp
PROBLEM SOLVING		Wetherall81 va			Wetherall81 va	
SIMPLE RT			Laurell78va Extrap		Laurell78va Extrap Lisper73va	
RANDOMIZE	Wetherall81 v-					

P-LEXICAL DECISION	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
CHOICE RT		Becker76va p?			Becker76va p?	
SIMPLE RT		Becker76va p?			Becker76va p?	

P-FLIGHT SIMULAT.	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
CHOICE RT	Bortolussi87 vvp Bortolussi86 vv				Bortolussi87 vvp Bortolussi86 vv	
MICHON	Wierwille85 (2,3,4)v-			Wierwille85 (3,4)v-	Wierwille85 (2)v-	
MONITOR				Soliday65vv		
TIME ESTIMATE	Bortolussi87 v-p Bortolussi86 v- Wierwille85(1,2)v- (3,4)				Bortolussi87 v-p Bortolussi86 v- Wierwille85 (1,2)v- (3,4)	
NAVIGATE	Lewis68vvp (field study)				Lewis68vvp (field study)	

P-IDENTIFY	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MEMORY		Klein76va			Allport72 a,v/ap	
MONITORING	Dornic80 (1)vap (2)vvp				Dornic80 (1)vap (2)vvp Chiles79vv	
IDENTIFY	Allport72 avp			Allport72 avp		
PSYCHO-MOTOR		Klein76vv				

P-SIMPLE RT	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING		Schmidt84av			Schmidt84av	
MENTAL MATH		Chiles70vv				
MICHON			Vroon73avp		Vroon73avp	

P-MEMORY	P-	P-DOWN	P-UP	S-	S-DOWN	C-UP
TRACKING		Johnston70 av			Johnston70 av	
CHOICE RT		Logan70 (1,2,3)avp			Logan70 (1,2,3)avp	
MEMORY		Broadbent62 va Chow75 (1,4)vv (3)va			Shulman71 (1)av	
MENTAL MATH		Silverstein71 vv Roediger75 (3)vv				
MICHON		Roediger75 (1,2)vv		Roediger75 (1,2)vv		
MONITOR	Tyler74va	Mitsuda65aa Lindsay69aa Chow75(2)va		Linday69aa		
PROBLEM SOLVING		Trumbo70vv			Trumbo70vv	
SIMPLE RT					Martin74 (1,2)avp	
RANDOMIZE		Trumbo70v-				
DETECTION		Hoffman83 vvp			Shulman71 (2)av Shulman71 av Hoffman83 vvp	
IDENTIFY		Mitsuda65aa				
CARD SORTING		Murdock65 (1)av Murdock65 (2)avs			Murdock65 (2)avp Murdock65 (1)av	
DISTRACT		Glanzer66 a+v,v				

P-MONITOR	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING	Griffiths71 av				Griffiths71 av	
CHOICE RT		Smith69 (1,2)av			Krol71vap Smith69 (1,2)av	
MEMORY	Moskowitz74 aa	Chechile79 v+a,v			Chechile79 v+a,v Moskowitz74 aaExtrap	
MENTAL MATH		Kahneman67 vap Chiles70vv			Kahneman67 vap Chiles70vv	
MICHON		Shingledecker 83v-		Shingledecker 83v-		
MONITOR		Fleishman65 vv Hohmuth70 (1)av/va Long76(1,2) va Stager71vas Chechile79 v+a,v Goldstein78 vv	McGrath65 a/v,a+v	Stager71vas	Long76(1,2) va Hohmuth70 (1,3)av/va Chechile79 v+a,v	
PROBLEM SOLVE	Gould67vvp Smith66vv					Smith66vv
DETECTION		Tichner72vv				
IDENTIFY					Savage78 vvp	
OCCLUSION	Gould67vvp					
TIME ESTIMATE		Liu87v-			Liu87v-	

P-PSYCHO-MOTOR	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MICHON		Michon64(2) v-(uses multiple task comparisons)			Michon64(2) v-(uses multiple task comparisons)	

P- TRACKING	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING	Mirchandani 72vvp	Hess74vv Wickens80vv Wickens81vv	Tsang84vvp		Tsang84vvp Wickens81vv	Mirchandani 72vvp
CHOICE RT		Trumbo72 (1,2)va Benson85vvp Wempe68va Israel80vap Girouard84 (1,2)va Klapp84 vsap=	Gopher77vv	Gopher77vv	Damos73vvp Israel80vap Benson85vvp Girouard84 (1,2)va Klapp84va sp=	
MEMORY	Finkelman54 va Zeitlin75va	Trumbo71 (1,2)vv Noble67va Helmstra70 va Wickens80va Huddleston71 vvp Wickens81va	Tsang84 v,v/ap	Trumbo71 (1,2)vv Noble67va	Zeitlin75va Finkelman54 va Helmstra70 va Tsang84 v,v/sp Wickens80va Huddleston71 vvp Wickens81va	
MENTAL MATH	Huddleston71 vvp	McLeod73va Wickens81 Bahrick54va Chiles70 Helmstra70 va		Bahrick54va Helmstra70 va	McLeod73va Wickens81 Huddleston71 vvp	Chiles70
MICHON		Shingle- decker83v-			Shingle- decker83v-	
MONITOR	Schori73vv Bell78vap Kyriokides 77vvp Gabriel68vv Huddleston71 vvp Kelly67vv Figarola66va	Bergeron 68vv Herman65aa Putz74vv Kramer84 v,v/a Helmstra 70vv Malmstrom 83va Monty65vv		Figarola66va Kramer84 v,v/a Malmstrom 83va	Schori73vv Bell78vap Herman65aa Bergeron68 vv Kyriokides 77vvp Putz74vv Gabriel68vv Huddleston 71vvp Kelly67vv Monty65vv	Helmstra70 vv
PROBLEM SOLVING		Trumbo67va Figarola66vv		Trumbo67va Figarola66vv		
SIMPLE RT		Helmstra70 vv Kelly85va Wickens77vv Klapp84va sp=			Wickens77vv	Helmstra70 vv

P- TRACKING	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
STERNBERG		Wickens85 vap	Briggs72(2) va		Briggs72 (1,2)va Wickens85 vap	
SPATIAL TRANSFOR.					Vidulich85 v,v/ap	
CLASSIFY		Wickens81vv			Wickens81vv	
RANDOMIZE	Zeitlin75	Truijens76 va(cues)		Zeitlin75	Truijens76 va(cues)	
DETECTION		Wickens81vv			Wickens81vv	
IDENTIFY		Gabay77vv		Gabay77vv		
PSYCHO- MOTOR		Bergeron68 Wickens76			Bergeron68	

P- PROBLEM SOLVE	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
TRACKING	Wright74(2) vvp				Wright72(2) vvp	
CHOICE RT	Fisher75avp Extrap				Fisher75avp Extrap	
MEMORY	Daniel65va	Stager72vv				
MICHON		Michon64(2) v- (multiple task comparisons)			Michon64(2) v- (multiple task comparisons)	
MONITOR	Wright74(1) vvp				Wright74(1) vvp Chiles79vv	
PROBLEM SOLVING					Chiles79	

P-SPATIAL TRANFORM	P	P-DOWN	P-UP	S-	S-DOWN	S-UP
IDENTIFY		Fournier76 vap			Fournier76 vap	

P-MENTAL MATH	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MEMORY	Mandler73 avp				Mandler73 avp	
MONITOR	Dornic80 v?vp				Dornic80 v?vp Chiles79vv	
MICHON		Michon64(2) v- (multiple task comparisons)			Michon64(2) v- (multiple task comparisons)	

P-STERN-BERG	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MICHON		Shingledecker 83v-		Shingledecker 83v-		

P-TAPPING	P-	P-DOWN	P-UP	S-	S-DOWN	S-UP
MENTAL MATH	Kantowitz76 (1,2)vap				Kantowitz76 (1,2)vap Kantowitz74 vvp	